AFOSR 67-0980

AD651630

ASSOCIATIVE ADJUSTMENTS TO REDUCE ERRORS IN DOCUMENT SCREENING

by

Edward C. Bryant
Donald T. Searls
Robert H. Shumway
David G. Weinman

Prepared for

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH

under contract

AF 49(638)-[redacted] 1671

March 31, 1967

[redacted]

WESTAT RESEARCH, INC.
7979 Old Georgetown Road
Bethesda, Maryland

Distribution of this document is unlimited.

20040831018

I. INTRODUCTION

The term "association" with respect to document storage, search, and retrieval is subject to many interpretations. When one indexes he associates index terms with documents, when he classifies he associates "like" documents, and when he searches he associates documents with the presumed need of the user. However, the word association in the documentation field has come to imply a mathematical or statistical means by which the user is led to specify a broader spectrum of documents than he would otherwise consider and which at the same time sharpens his attention to those documents of highest presumed interest. The mathematical examination and evaluation of various proposed procedures, some classical and some new, has been the principal objective of this study. Experimental results based on a small sized data file and a few searches are presented to illuminate the ideas behind some of the newer methods.

II. ASSOCIATION AND ERROR

We have considered the use of associative techniques for the accomplishment of one principal objective which is the assignment of a "relevance number" to each document in the file, the relevance numbers presumably reflecting the searcher's interest in each document. There are several kinds of errors occurring in this assignment procedure which [redacted] guarantee that the ordered ranking generated does not adequately represent the ranking which the searcher would prefer if he examined exhaustively every document in the file.

Generally, in a retrieval system, the documents in the file are indexed with a number of terms so that each document in the file is arbitrarily specified by an index vector of zeros and ones, say $\underline{c}_i' = (c_{i1}, c_{i2}, \ldots, c_{it})$ where $c_{ij} = 1$ if term j is relevant to document i and zero otherwise. The first kind of error which can occur is in the indexing of the vector $\underline{c}_i'$ and we have referred to these errors in the past as overindexing (assign $c_{ij} = 1$ when $c_{ij} = 0$) and underindexing (assign $c_{ij} = 0$ when $c_{ij} = 1$) respectively. The influence of underindexing has been reduced by means of adjustment procedures which change the original coding c_{ij} to some new value b_{ij}. Two papers, on different adjustment procedures, are included in the Appendix.

The simplest query is composed of zeros and ones and "searching" consists of a process of matching queries with the indexed terms for documents. In one of its simpler forms, one accumulates the number of terms for which the match is perfect, using this as a retrieval score for the document. A modification is to specify a subset of terms over which the match is to be performed, the remaining terms being presumed to have no bearing on the search objective. Hence, a second kind of error may arise from the searcher's inability to formulate a query $\underline{q}' = (q_1, q_2, \ldots, q_t)$ correctly, at least from a point of view which returns the relevant references. In this study query adjustment may be treated as the dual of file adjustment, with an additional complicating factor posed by the possibility of augmenting or reducing the number of terms in the query.

Some experimental results will be presented which indicate that augmenting the basic query by purely mechanical means is not necessarily effective.

To summarize, we may consider errors as arising from either indexing errors or searching errors with the file-oriented adjustments correcting indexing errors and the query adjustments correcting errors in query formulation.

III. SOME PROPOSED ASSOCIATIVE METHODS

a. Notation

In the following sections the original file consisting of d documents coded with t terms will be represented as a dxt matrix of zeros and ones, say $\left\{c_{ij}\right\}$, i=1, 2, . . . , d, j=1, 2, . . . , t, with $c_{ij} = 1$ if the jth term is relevant to the ith document and $c_{ij} = 0$ otherwise. A single query will be defined as a vector of ones and zeros $\underline{q}' = (q_1, q_2, \ldots, q_t)$ with a 1 in the query indicating an interest in the presence of a term and a zero indicating an interest in its absence. If a term is of no interest it is not included in the query, so that the dimension of $\underline{q}'$ is reduced accordingly to s $(s < t)$, $(\underline{q}' = (q_1, \ldots, q_s))$. If an adjustment is made to the file this adjustment is denoted by $c^* = (c^*_{ij})$ i = 1, . . . , d, j = 1, . . . , t. Adjustments to the query are denoted by $\underline{q}^{*'} = (q^*_1, \ldots, q^*_t)$.

b. Linear Associative Retrieval

A pioneering extension of the coordinate indexing scheme was proposed by Giuliano and Jones [1] in which the following linear model was postulated:

(a) The ranking r_i of the ith document is a linear combination of the terms contained in the adjusted query q^*_j and the adjusted codings.

$$r_i = \sum_{j=1}^{t} c^*_{ij}\, q^*_j \qquad i = 1, \ldots, d \tag{1}$$

(b) The adjusted value q^*_j is the sum of the original value q_j in a query and a linear combination of the adjusted codings for the documents containing it.

$$q^*_j = q_j + \lambda_j \sum_{i=1}^{d} r_i\, c^{**}_{ij} \tag{2}$$

where the λ_j are weighting factors for the adjustment to the jth term in the query and c^*_{ij} and c^{**}_{ij} are normalized matrices derived by dividing each row of C by its sum and each row of C' by its sum. If Λ is a txt diagonal matrix made up of $\lambda_1, \ldots, \lambda_t$ then equations (1) and (2) can be rewritten:

$$\underline{r} = \underline{C}^*\, \underline{q}^* \tag{1'}$$

$$\underline{q}^* = \underline{q} + \Lambda C^{**}\underline{r} \tag{2'}$$

which have the solution:

$$\underline{r} = C^*(I - \Lambda C^{**}C^*)^{-1}\underline{q} = C^*K\underline{q} \tag{3}$$

In this case we see that the matrix K can be regarded as a right hand transformation of the original normalized coding matrix or as a left hand transformation operating on the original query. Hence, it may be regarded as either adjusting for errors in the original coding or adjusting the query or both. One may interpret the parameter λ_j as being assigned on the basis of the confidence of the searcher in the value q_j of his original query. If he is relatively confident that q_j should be near q_j^*, see equations (1) and (2), a small weight λ_j would be assigned as a correction to q_j. If he is less certain about some terms than others, he uses the λ_j to convey larger weight to the associative corrections that the system assigns these terms. We note that the relevance measure proposed in (3) is the dot product of two vectors at least one of which has been transformed. One can easily show that this response need not be as indicative of the closeness of match between the two vectors as certain other matching operations, which we shall discuss in later sections.

c. Probabilistic Indexing

An indexing and retrieval scheme which seems to have a great deal of merit is the probabilistic indexing idea of Maron and Kuhns [2]. We shall sketch their argument here and discuss a parallel extension.

Maron and Kuhns argue that the index value c_{ij} should be an estimate of the probability $P(j|R_i)$ that a user will request term j given that he is interested in document i (or that document i would be relevant). The authors further propose a method for refining the estimates c_{ij}, so that, for convenience in notation, it will be assumed that the actual values $P(j|R_i)$ are known. The prior probability $P_o(R_i)$ of relevance of document i will also be assumed given. (If no prior information about document i is available, $P_o(R_i)$ can be chosen to be any number, fixed for all i.) Finally, the probability P(j) that a user will request term j is assumed known. It could be found from library statistics. It follows that the probability $P(R_i | j)$ of relevance of document i, given that term j is requested, is:

$$P(R_i | j) = \frac{P(j | R_i) P_o(R_i)}{P(j)} \tag{4}$$

The numbers $P(R_i | j)$ for all documents i would be sufficient to rank the documents in order of probable relevance if term j were the only term requested. This is seldom the case, however. Maron and Kuhns next consider the case of Boolean queries. Let "$j \wedge k$" signify that documents with *both* terms j and k are requested, and let "$j \vee k$" signify that documents with *either* term are requested. From elementary rules of probability, it follows that:

(i) $0 \leq P(j \wedge k | R_i) \leq 1, \quad 0 \leq P(j \vee k | R_i) \leq 1$

(ii) $P(j \wedge k | R_i) \leq P(j | R_i)$

(iii) $P(j \vee k | R_i) = P(j | R_i) + P(k | R_i) - P(j \wedge k | R_i)$

(iv) $\max \left[0, P(j | R_i) + P(k | R_i) - 1 \right] \leq P(j \wedge k | R_i) \leq \min \left[P(j | R_i), P(k | R_i) \right]$

If $P(j \wedge k | R_i)$ were known, then $P(j \vee k | R_i)$ could be computed from (iii). Although $P(j \wedge k | R_i)$ is not known, inequalities (iv) serve to bound it above and below. The authors suggest using for $P(j \wedge k | R_i)$ the "independence value" $P(j | R_i) P(k | R_i)$, which always lies between the bounds in (iv). With this convention for conjunctions, the probability $P(Q | R_i)$ of any query composed of conjunctions and disjunctions of index terms can be computed from the individual values $P(j|R_i)$ of the terms j in the query. It then follows, as in (4), that the posterior probability $P(R_i|Q)$ of relevance of document i given the query Q is:

$$P(R_i|Q) = \frac{P(Q|R_i)P_o(R_i)}{P(Q)} \qquad (5)$$

Although the probability P(Q) that query Q will be used cannot generally be known, it may be assigned an arbitrary value for a given search. Thus, documents ranked by scores:

$$P(R_i|Q)P(Q) = P(Q|R_i)P_o(R_i)$$

will be ranked in the same order as documents ranked by (5), regardless of the value of P(Q).

It should be noted that the authors begin with a probabilistic indexing. Most of the files we are concerned with have a coordinate (0, 1) indexing which we use as the basis for some type of transformation. In what follows we develop a probabilistic model to fit such files.

Although many existing files have (0, 1) indexing (or some other non-probabilistic indexing), it should be possible to estimate a probability such as $P(j|R_i)$, the probability that term j will be indexed in a randomly selected relevant document. Reasoning analogous to that of Maron and Kuhns should then lead to an estimate of $P(R_i)$, the probability of relevance of document i. Our notation will resemble that of Maron and Kuhns. However, whenever a term symbol, j, k, m, appears, it refers to the event that term j, k, or m is indexed. In the previous section, it referred to the event that term j, k, or m was requested. Definitions of the expressions in this section are as follows:

1. $P(j|R_i)$ is the probability that term j is indexed for document i, given that document i is relevant. $P(j|\bar{R}_i)$, is the analogous probability, given that document i is irrelevant. If the number of relevant documents is small, this is estimated by the frequency of indexing of j.

2. $P(R_i|j)$ is the probability that document i is relevant given that term j is indexed.

3. $P_o(R_i)$ is the prior probability of relevance of document i (the same as in the Maron and Kuhns method). $P_o(\bar{R}_i) = 1-P_o(R_i)$ is the prior probability of irrelevance of document i.

4. $P(R_i)$ is the final relevance score of document i.

The first problem is that of estimating $P(j|R_i)$. Suppose a large number of coordinate searches have been made and, for each query, the indexings of relevant documents have been listed. For a given term k, consider all queries containing k and the indexings of all documents relevant to each of these queries. For any term j, let $p_{k,j}$ be the relative frequency with which term j appears in this set of indexings. For example $p_{k,k}$ is the proportion of times term k appeared in relevant documents when term k was requested. The values $p_{k,k}$ then could be used as an estimate of $P(k|R_i)$ whenever term k is requested. It is possible to make use of all terms however, not just those which are requested. If term j is not requested,

$P(j|R_i)$ can be estimated by $(1/n)\sum_k p_{k,j}$, summed over all terms k in the query, where n is the number of terms in the query. (This is an assumption which should be verified. We did not have sufficient data to test it.)

Applying Bayes' Rule, we obtain:

$$P(R_i|j) = \frac{P(j|R_i)P_o(R_i)}{P(j|R_i)P_o(R_i) + P(j|R_i)P_o(R_i)} \tag{6}$$

For document i, we obtain $P(R_i|j)$ for all terms j indexed in document i. We now have the problem of computing $P(R_i)$ from the values $P(R_i|j)$. As with Maron and Kuhns method, there is no obvious formula for the computation.

If the events (relevance and term j indexed) are <u>independent</u> for all j, we would have:

$$P(R_i) = \sum_j P(R_i|j) - \sum_{j<k} P(R_i|j)\,P(R_i|k) + \ldots \pm P(R_i|j)$$

$$\pm \sum_{j<\ldots<m} P(R_i|j)\ldots P(R_i|m) \tag{7}$$

If $P_o(R_i)$ is small, the values $P(R_i|j)$ will also be small, so that (7) would reduce to:

$$P(R_i) = \sum_j P(R_i|j) \tag{8}$$

where the sum is over all terms j indexed for document i.

The data we have indicates that these methods of ranking documents give high scores to documents with many terms indexed, even if the terms were not particularly desirable. Formula (6) shows that $P(R_i|j)$ is a posterior probability whose prior probability is $P_o(R_i)$. It would seem that a term indicates relevance if $P(R_i|j) > P_o(R_i)$, and otherwise indicates irrelevance. Formulas (7) and (8), however, give a document a high score if there are many terms j with $P(R_i|j) < P_o(R_i)$, that is, many terms indicating irrelevance. This difficulty can be avoided by defining.

$$P^*(R_i) = \sum_j \left[P(R_i|j) - P_o(R_i) \right] \tag{9}$$

where the sum extends over all terms j indexed for document i. It is obvious that $P^*(R_i)$ will not be a probability, since the values $P(R_i|j)-P_o(R_i)$ are not probabilities. These values can, however, be thought of as weights attached to terms. Once a query is formulated, weights $P(R_i|j)-P_o(R_i)$ can be computed for each term in the file. The score $P^*(R_i)$ of document i is then computed as the sum of the weights of the terms in the indexing of document i.

d. Measures of Mismatch

In this study it has seemed appropriate to define a measure of distance between a document in the file and the query which takes more factors into account than the simple matching of terms when they are present. For example, with zero-one indexing and zero-one queries the simple dot product

$$r_i = \sum_{j=1}^{t} c_{ij} q_j \qquad i = 1, \ldots, d \tag{10}$$

method of matching the documents with the query increases relevance r_i

only when $c_{ij} = q_j = 1$. However, one may also want to be able to decrease the measure of relevance for terms that would definitely never be found in a relevant document. ·The dot product assigns these terms a zero contribution in the relevance measure. The effect is the same as if one had not cared whether the term was present or not.

A class of distance measures which can decrease the relevance for terms whose absence is desired are the mismatch measures. The unweighted mismatch for ith document is defined as

$$M_i = \sum_{k=1}^{t} (c_{ik} - q_k)^2 \tag{11}$$

where the c_{ij} is the original document coding for the jth term in the ith document and q_k is the query indexing for the kth term. The above measure may be computed for $i = 1, 2, \ldots, d$ with the ordered M_i's constituting a ranking for the documents relative to the query $\underline{q}$. Note that the documents with the lowest mismatch are those of highest presumed relevance. The squares are retained in (11) to allow for the possibility of having the adjusted codings or queries defined on a continuous scale. In this case

$$M_i^* = \sum_{k=1}^{t} (c_{ik}^* - q_k^*)^2 \tag{12}$$

Hence, for the mismatch type operators the user may specify a subset of the total terms which, if indexed, indicate his interest in the document. He may also specify a set of terms which indicate his lack of interest. The former appear in the search query as ones and the latter as zeros (or some continuous approximation thereof). He is presumed to be indifferent to the remainder of the possible terms so they should not appear in the matching operation. A previous report [3] discusses methods for adjusting the file using a regression estimate for c^*_{ij}. (See summary of this technique in the Appendix to this report).

In order to generalize the above measure of mismatch, we let $\underline{c}'_i = (c_{i1}, c_{i2}, \ldots, c_{it})$ be the vector of codings for the ith document with $\underline{q}' = (q_1, q_2, \ldots, q_t)$ the vector of codings for the query. Now if $W = \{w_{ij}\}$ with i, j = 1, 2, ..., t is a generalized weight matrix we may define a generalized mismatch between the ith document and the query as

$$M_i = (\underline{c}_i - \underline{q})' W(\underline{c}_i - \underline{q}) \tag{13}$$

For the case W = I, the above is written

$$M_i = \sum_{k=1}^{t} (c_{ik} - q_k)^2 \tag{14}$$

which is just the ordinary mismatch. If the adjusted codings and the queries are zeros or ones then with W = I equation (13) just becomes the number of mismatched terms. Hence, equation (13) is an appropriate generalization in the sense that the special case W = I leads to the ordinary measure of mismatch. The entries w_{ij} in the weighting matrix w_{mn} weight the mismatches $(c_{im} - q_m)'(c_{in} - q_n)$ proportionately to the term indexes m and n. For example, for m ≠ n the weight w_{mn} determines the extent to which that mismatch contributes to the total mismatch M. The statistical properties of interest for this generalized mismatch relate to its behavior for relevant and irrelevant documents. We shall examine here the mean and variance of the mismatch for a general weighting matrix W. For a particular choice of W, the probability distribution of M is well known both for relevant and irrelevant documents. Hence, we may derive the missed document and false retrieval rates as well as a procedure for setting a cutoff point for the mismatch which determines a prespecified missed document rate.

In order to examine the behavior of the mismatch we must make some assumptions about the distributional form of the multivariate vector $\underline{c}$ of codings or adjusted codings. Suppose that $\underline{c}$ is a multivariate normal random vector with a mean depending on whether $\underline{c}$ is a sample from a population of relevant documents or a sample from a population of irrelevant documents and with a constant covariance matrix. That is, we have the summary formulae given in Table 1 below.

Table 1. Mean and Covariance of Document Coding Vector $\underline{c}$

	Mean	Covariance
Relevant	$E_R\underline{c} = \underline{q}$	$E_R(\underline{c} - \underline{q})(\underline{c} - \underline{q})' = \Sigma$
Irrelevant	$E_I\underline{c} = \underline{q} + \underline{\epsilon}$	$E_I(\underline{c} - \underline{q} - \underline{\epsilon})(\underline{c} - \underline{q} - \underline{\epsilon})' = \Sigma$

We wish to examine the statistical behavior of a generalized mismatch defined by

$$M = (\underline{c} - \underline{q})' W(\underline{c} - \underline{q}) \qquad (15)$$

for relevant and irrelevant documents. A desired property of M is that the expected mismatch should be low for relevant documents and high for irrelevant documents with a small variance. Figure 1 shows how a large separation with relatively small variance enhances the discriminatory capabilities of the generalized mismatch. If we decide to examine every document whose generalized mismatch is less than a fixed cutoff point K, then the shaded areas represent the resulting missed document and false retrieval rates. If we know the exact probability distributions these error rates can be specified in advance by choosing the proper cutoff point. It must be noted in passing that this processing scheme gives neither minimum total error probability nor is it optimum in the Neyman-Pearson sense. More will be said about this later.

If the assumptions in Table 1 are met we may derive the expectation and variance of the mismatch operator for several different weighting schemes. (For the derivation for a general weighting matrix W see Appendix.) The variance computations require the $\underline{c}$'s to have a joint multivariate normal distribution and we note that for the unadjusted codings this would not be the case. Table 2 shows the results for a general W and also for the choices W = I (eq. 14) and W = Σ^{-1}.

The most interesting case seems to be the mismatch

$$M = (c - q)' \Sigma^{-1} (c - q) \tag{16}$$

The distribution of M is chi square with n degrees of freedom for relevant documents and follows a noncentral χ^2 distribution with noncentrality parameter $\underline{\epsilon}' \Sigma^{-1} \underline{\epsilon}$ for irrelevant documents.

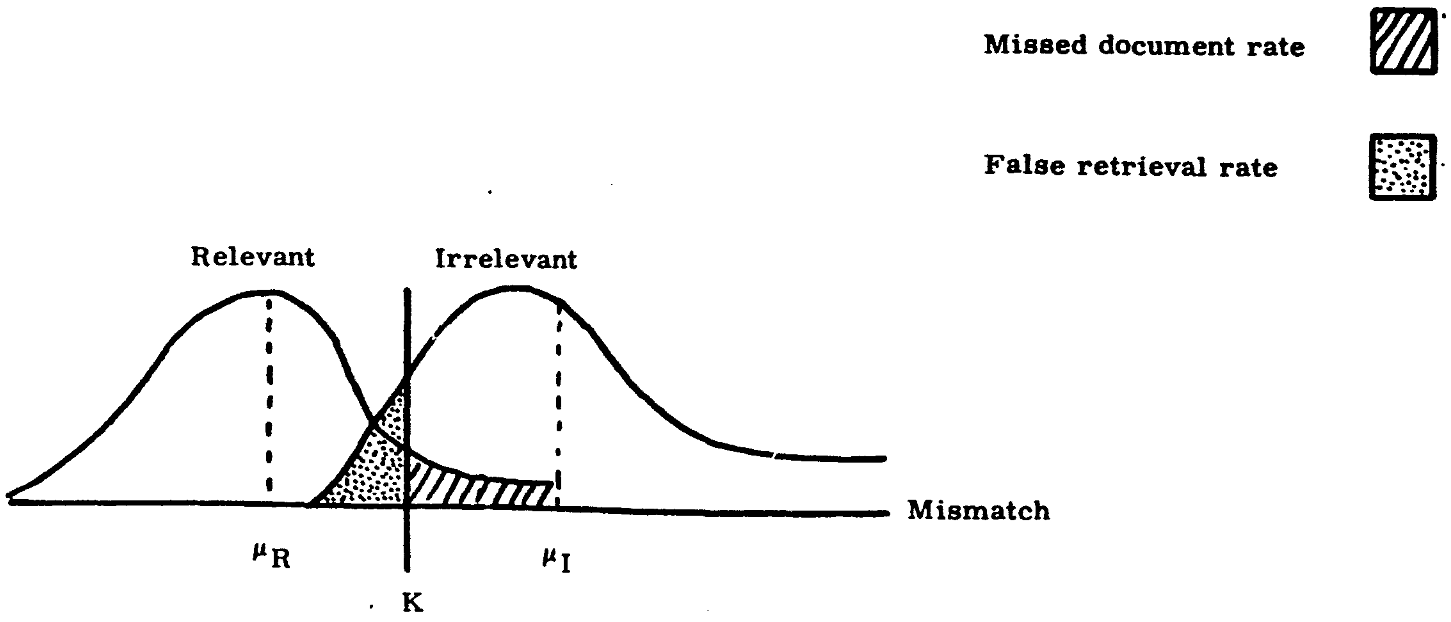

Figure 1. Hypothetical distribution of the generalized mismatch for relevant and irrelevant documents

Table 2. Mean and Variance of the Generalized Mismatch

	Mean	Variance
$W = W$		
Relevant	$\mathrm{tr}\ W\Sigma$	$2\mathrm{tr}\ (W\Sigma)^2$
Irrelevant	$\mathrm{tr}\ W\Sigma + \underline{\epsilon}'W\underline{\epsilon}$	$2\mathrm{tr}\ (W\Sigma)^2 + 4\underline{\epsilon}'W\Sigma\ W\underline{\epsilon}$
$W = I$		
Relevant	$\sum_i \sigma_{ii}$	$2\sum_{ij} \sigma_{ij}^2$
Irrelevant	$\sum_i (\sigma_{ii} + \epsilon_i^2)$	$2\sum_{ij} (\sigma_{ij}^2 + 4\epsilon_i\sigma_{ij}\epsilon_j)$
$W = \Sigma^{-1}$		
Relevant	n	$2n$
*Irrelevant	$n + \sum_{ij} \epsilon_i \sigma^{ij} \epsilon_j$	$2n + 4\sum_{ij} \epsilon_i \sigma^{ij} \epsilon_j$

* $\left\{\sigma^{ij}\right\} = \Sigma^{-1}$

Hence, for an n term search one can determine a cutoff point by consulting the α% point for the χ^2 distribution. This determines a search with a missed document rate of α. For example, with a 10 term search and a desired missed document rate of .05, $\chi^2_{.05} = 18.31$. Hence, the searcher would examine only those documents whose generalized mismatch was less than 18.31. The false retrieval rate would depend on the non-centrality parameter $\underline{\epsilon}' \Sigma^{-1} \underline{\epsilon}$, a characteristic of the query and the file. Thus, one could develop, under the appropriate restrictions, a theory which could guarantee certain error rates if the weighting matrix is taken to be $W = \Sigma^{-1}$.

Cooper [4] has discussed the conditions under which quadratic classification functions of the form

$$M_R = (\underline{c} - \underline{q})' W_R (\underline{c} - \underline{q}) + E_R \tag{17}$$

$$M_I = (\underline{c} - \underline{q} - \underline{\epsilon})' W_I (\underline{c} - \underline{q} - \underline{\epsilon}) + E_I \tag{18}$$

are optimal with E_R and E_I constants to be determined. In this case one computes M_R and M_I for each document and then assigns it to the class of relevant documents if $M_R \leq M_I$. In general the principle of optimality satisfied is that the total probability (missed document rate and false retrieval rate) of misclassification is a minimum when the probabilities of relevance and irrelevance are equal. It is "Neyman-Pearson" also which means that a minimum false retrieval is achieved for a fixed missed document rate. Cooper shows that the optimal weighting matrices for $\underline{c}$, a multivariate normal, Pearson Type II, or Pearson Type VII distribution, are given by Σ^{-1} (Σ_R^{-1}, Σ_I^{-1}) where Σ_R and Σ_I are the covariance matrices for relevant and irrelevant documents.

The basic difficulties associated with applying mismatch measures of the forms (13) and (14) occur because of the difficulty in estimating the vector $\underline{\epsilon}$ which is the difference between the means of the relevant and irrelevant documents. Also, in practical cases, there is a problem with the handling of terms in the query toward which the searcher is totally indifferent. More will be said about this matter in a following section.

IV. FILE ADJUSTMENT PROCEDURES

In Section II it was pointed out that a principal source of error in retrieval systems is indexing. Since the presence in a document of the concepts represented by index terms tends to be correlated, one is led to believe that by associative means he can improve upon the assignment of index terms to documents. An important part of the research under this and a previous contract [3] has explored this possibility. Two technical papers covering this research appear in the appendix to this report:

The first paper gives the conditions under which one would expect to gain (in a well defined sense) by replacing ones and zeros in the proxy file by numbers which depend upon interrelationships among the index terms, indexing frequency, and the actual 0-1 indexing assigned by the system.

If one can ignore errors of over indexing and if queries are formed so as always to search for the presence of a term, rather tha₁ its absence, one can expect gains under quite general conditions. These conditions may be expected to hold in a large number of real life files. Let c_{ij} denote the 0-1 indexing (coding) of the jth term with respect to the ith document. Then, one will expect to gain by file adjustment whenever $c_{ij} = 0$ can be replaced by a value b_{ij} which, on the average, is greater when the jth term should be indexed than when it shouldn't.

To illustrate, designate c_{ij} as the "original indexing" and b_{ij} as an "imputed indexing" which is found by using any predictive information available. Conceptually, there is a "correct" indexing which is either zero or one. If the imputed indexing is, say, 0.70 when the correct indexing is one and 0.20 when the correct indexing is zero, there should be, on the average, a net gain in replacing the actual indexing (of zero) by the imputed indexing.

If the searcher sometimes specifies the absence of the jth term as a condition for relevance, the situation is not so clear, and some graphs are included to show when one can expect to gain under these circumstances.

No generalizations have been drawn for cases in which overindexing is substantial. Also, generalizations have not been found for situations in which it is impossible, because of the nature of the subject matter, to specify what is a "correct" indexing. However, some experimentation suggests that file adjustment may prove useful under these circumstances as well.

The second paper in the Appendix considers the same general circumstances as the first paper, but assumes that the file adjustment procedure will replace a zero indexing with a one, rather than with some other imputed indexing which can have any value, but which usually lies between zero and one. Procedurally, one arrives at an indicator variable, which might be the same as the imputed indexing in the first paper. If the indicator variable is greater than a prescribed cutoff, a one is assigned. Otherwise the indexing remains at zero. The cutoff can be adjusted so as to minimize total error, or some value function of the two types of classification error. Some graphs which show the probability of positive gain are included.

V. EXPERIMENTAL RESULTS

A small file of electronics patents from the United States Patent Office was chosen for an experiment comparing some of the methods proposed in Section III. The patents disclose analog to digital and digital to analog converter features. An example of the indexing sheet is given in reference [3]. A subset of 478 documents having the analog to digital feature was chosen for the experiment. A group of 13 searches for which complete data were available, including the terms in the query and the identity of the relevant documents in the base sample, was also selected. The original term list was reduced to a subgroup of 38 terms which were either the principal terms in the 13 searches or were highly correlated with them.

Table 3 shows summary information on the searches with the exact number and identity of the relevant documents having been determined by an exhaustive search through the file. Table 3 also shows the ranking of the relevant documents for four different methods of searching.

Method (1) searches used the original file codings and query codings with the ranking determined from the mismatch measure (eq. 11). This method gives a baseline against which to compare the other proposed methods. Note that the possibility of tied values of the mismatch measure introduces an ambiguity into the ranking. This was resolved by using the expected value of the number of documents examined in the group of documents having the highest tied ranking. In general, if there are k wanted documents in a set of N with tied ranks and if one examines the N randomly, the expected number of examinations required to obtain the kth document is k(N+1)/k+1. Method (2) is a file adjustment procedure which computes a weighted estimate for the adjusted coding from the original coded value, a marginal estimate and a regression estimate [3]. Again, the ordinary mismatch is based on equation (12) with the adjusted codings c^*_{ij} in the mismatch.

The weighted mismatch in method (3) is based on equation (13) applied to the reduced file of 38 terms. Since the searches were not originally specified over all 38 terms, values of the query were filled in by assuming that if one were indifferent as to whether a term was present or not the value of the query q_j for that term could be replaced by the sample average

Table 3. Summary of Ranks of Wanted Documents in Experimental Searches

Search Number	Wanted Documents	Search Procedure * (1)	(2)	(3)	(4)
1	2975409	10.5	2	70	3.5
	3024990	10.5	13	22	18.5
2	2817704	1.5	2	1	82.5
	3049701	5.5	3	378	3.5
3	3051943	60.0	22	21	137.5
4	2715724	32.0	34	171	43.5
	2950469	32.0	19	379	59.0
	2974315	32.0	50	211	11.0
5	3066286	4.5	1	1	16.5
6	2612550	5.0	3	3	27.0
	2869079	31.5	10	393	4.5
	2950469	31.5	19	354	27.0
	3023405	31.5	31	38	27.0
	3030614	5.0	4	107	4.5
	3041469	31.5	13	281	27.0
	3050713	31.5	12	2	71.5
7	2931023	81.0	53	7	49.5
	2938198	81.0	30	4	49.5
	2938199	81.0	30	5	49.5
	3066286	13.5	3	8	8.5
8	2873440	10.5	2	277	55.0
	2873442	10.5	3	158	158.0
	2976528	10.5	11	467	211.5
9	2793360	49.0	29	177	336.5
	2928053	49.0	55	220	168.5
	2991462	3.5	2	212	80.0
10	3045230	2.0	2	10	134.5

Table 3. continued

Search Number	Wanted Documents	Search Procedure *			
		(1)	(2)	(3)	(4)
11	2436178	66.5	41	12	76.5
	2793807	16.5	6	52	40.0
12	2775755	15.5	30	268	31.0
	2901170	15.5	17	2	31.0
13	3051941	6.5	11	9	225.0

* The following procedures were used:

(1) Variable number of terms with no file adjustment

(2) Variable number of terms with regression file adjustment

(3) Thirty-eight terms with quadratic mismatch

(4) Variable number of terms with a correlation-adjusted query

$\sum_i c_{ij}/ d$ of that term over all the documents. The weighting matrix W was taken to be the inverse of the full 38 term variance covariance matrix.

Method (4) considered adjusting the query by using the correlations between the original terms specified in the search and other terms in the full group to augment the query. The assumption was made that, with high relative frequency, a term whose correlation with a term already in the query was greater than 0.30 should be present in relevant documents, whereas a term whose negative correlation with any term in the original query was less than -0.20 should be absent in relevant documents.

Hence, the original queries were augmented by asking for the presence of terms highly positively correlated with terms already in the query and by asking for absence of terms highly negatively correlated with terms already in the search query. Figure 2 shows a frequency distribution of values for intercorrelations among the 38 terms. On the basis of this distribution the cutoff values of -0.20 and +0.30 were established arbitrarily to serve as high negative and high positive correlations.

Method (5) serves for comparison of all the schemes by assuming that the relevant documents are uniformly distributed over the set of retrieved ranked documents. This method is equivalent to random searching.

A graphic way of comparing the results of the five methodo is given by the usual relevance profile (see Figure 3) which examines the proportion of desired references retrieved as one moves down through the list of relevant documents. We will consider here the performance measured for the 32

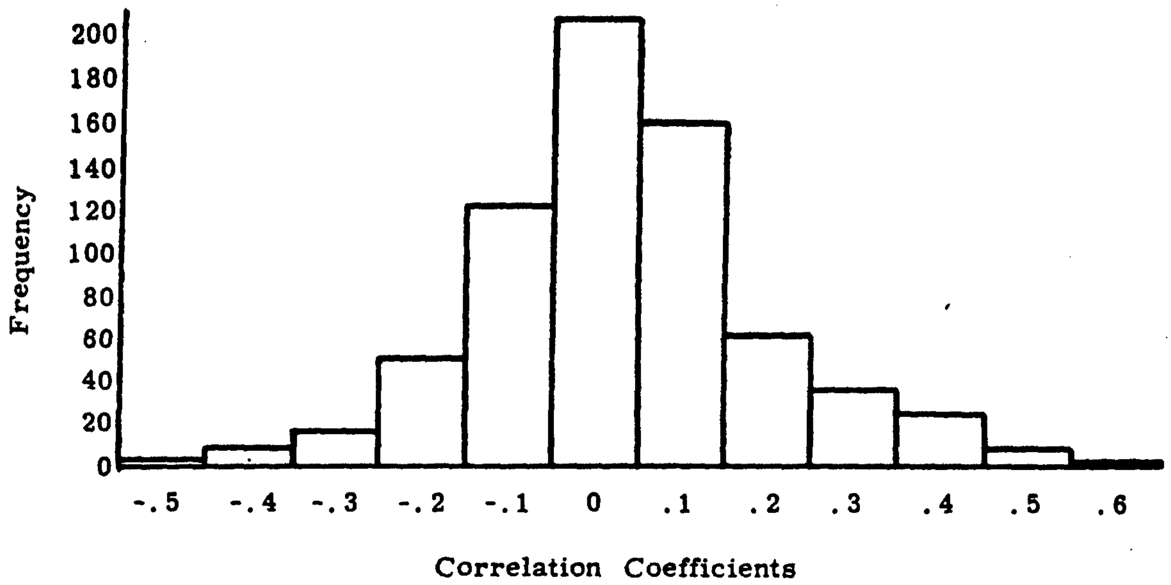

Figure 2. Simple correlations among indexing terms

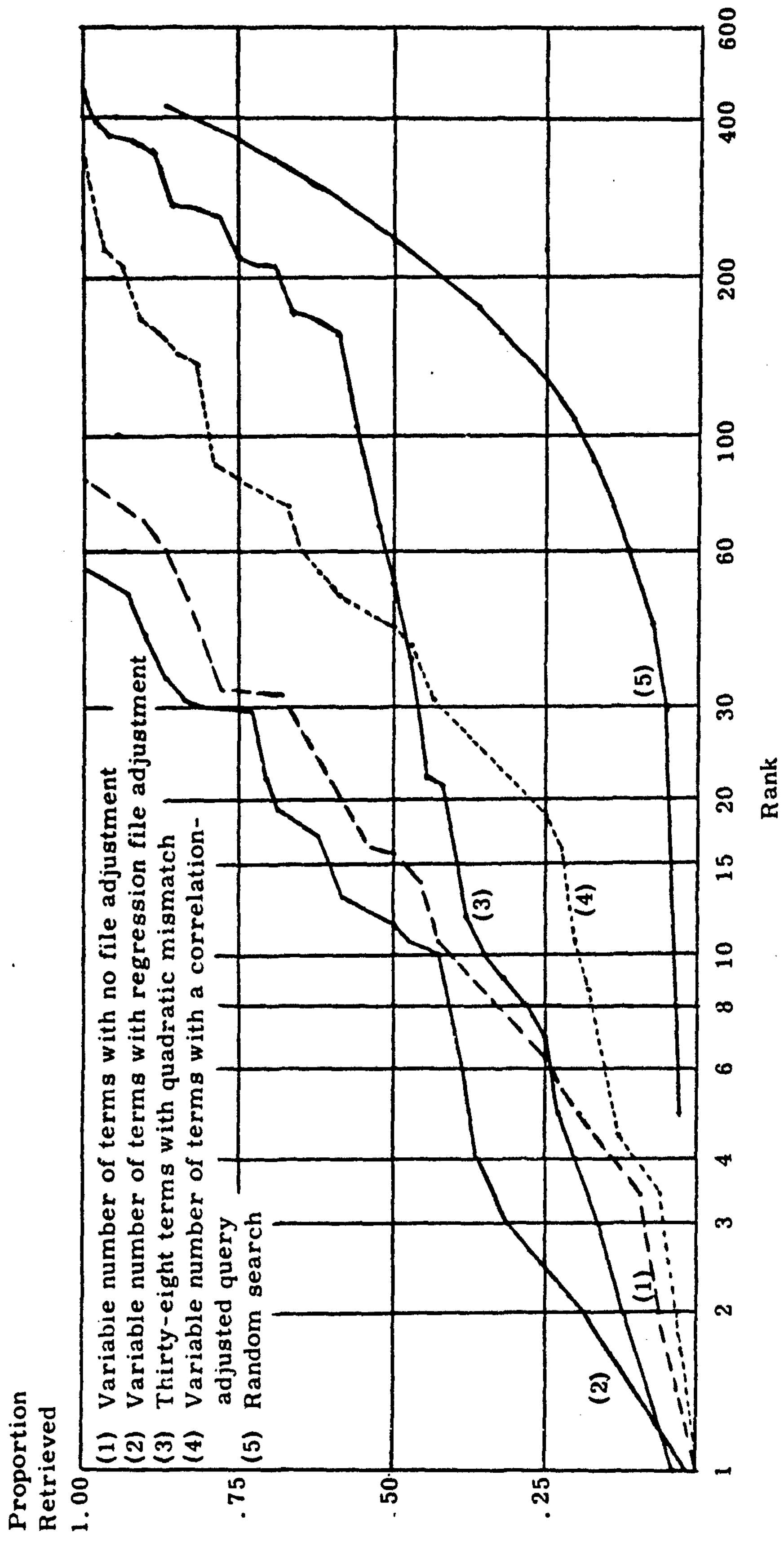

Figure 3. Retrieval Profil^s for Four Associative Methods

relevant documents over all of the 13 searches. Method (2) seems to give uniformly highest results with a particular superiority demonstrated at 3 documents where it yields twice as many relevant documents as the nearest competitor. The ordinary mismatch (method (1)) seems to begin to increase in effectiveness after five documents are examined and almost catches up with the file adjusted mismatch by the time ten documents have been examined. The augmented searches (method (4)) do not seem to perform very well over any part of the range while the quadratic mismatch (method (3)) gives results that are generally inferior to the regression adjusted file (method (2)). This is somewhat disappointing since the quadratic mismatch cah be related to certain optimal classification procedures. A possible source of error is the procedure of assigning query values for unspecified terms equal to their mean in the file.

The poor performance of the quadratic mismatch is not due to reducing the term list from 88 to 38. Only original searches 2, 8, 9, and 10 contained terms which were not in the reduced set of 38 terms. Eliminating these searches, the average rank under file adjustment (2) is 19.7 while quadratic mismatch yields 105.2. The query augmentation procedure (method (4)) is totally ineffective.

In summary, from the very small data sample examined, it would seem that associative file adjustments lead to substantial gains, with the usefulness of the more general mismatch measures not conclusively demonstrated.

VI. SUMMARY

This research has examined the sources of errors in the process of matching queries and indexed documents. It is known that both indexing and query formulation are subject to error, and the objective of associative adjustment is to minimize (in some general sense) the effect of such errors. One can adjust the indexings of the file and match unadjusted queries against them, he can match the unadjusted file against adjusted queries, or he can adjust both the file and the queries prior to performing the match.

On purely theoretical grounds it is difficult to choose between file adjustment and query adjustment, but on practical grounds file adjustment has the following points in its favor:

1. File adjustment need be done only once and can be done by computer on second-shift time, while query adjustment must be done at the time of the search.

2. Information by which the file can be adjusted is easier to obtain than information by which the query can be adjusted. Repeated indexings by randomly selected indexers (or even relationships among once-indexed terms) provides information by which the file can be adjusted. However, a searcher does a limited number of searches at best, and throughout the interval during which his searching is being done he is constantly changing his behavior -- perhaps adjusting to inadequacies of the file. Thus, an adjustment procedure which might be optimum for him at one time would be detrimental to his success at another.

Intuitively, there is merit in adjusting the file and letting it remain fixed so that the searcher may adjust his behavior to it. Our experimentation, admittedly fragmentary, has not revealed a search adjustment procedure which is better in any sense than file adjustment.

There is need for carefully designed and implemented research in operating files to develop empirically optimal file adjustment procedures. Holding these files fixed, it should then be possible for the searchers to optimize their searches.

In this investigation a theoretical measure of closeness or distance has been proposed which is a direct generalization of the notion of matching. This measure of mismatch enables one to specify terms which are definitely not to be present in a given relevant document as well as those which are. Theoretical properties of the mismatch measure which have been examined relate to two areas. The first is the influence of various file adjustment procedures on the mismatch measure, that is, the theoretical amount that one could expect to gain in employing an adjustment procedure. The second is the investigation of the statistical properties of the generalized mismatch including the prediction of the missed document and false retrieval rates. The experimental results based on a limited file size do not confirm the superiority of the generalized mismatch.

VII. REFERENCES

1. Vincent E. Giuliano and Paul E. Jones, "Linear Associative Information Retrieval," Chapter 2 of Howerton and Weeks, Vistas in Information Handling, Vol. 1, Spartan Books, Washington, D. C., 1963.

2. M. E. Maron and J. L. Kuhns, "On relevance, probabilistic indexing, and information retrieval," JACM, Vol. 7, No. 3, July, 1960.

3. Edward C. Bryant, Donald T. Searls and Robert H. Shumway, "Some theoretical aspects of the improvement of document screening by associative transformations", report to AFOSP under contract AF 49(638)-1484, Westat Research, Inc., Denver, November 30, 1965 (AD 628 191).

4. Paul W. Cooper, "Statistical classification with quadratic forms", Biometrika, Vol. 50, Nos. 3 and 4, 1963, pp. 439-447.

APPENDIX 1

ASSOCIATIVE CORRECTION FOR UNDERINDEXING*

by

Edward C. Bryant
Donald T. Searls
Robert H. Shumway

With respect to document storage and retrieval, one can think of associative techniques as either those which improve the file or those which improve the search query. The objective of both is to improve the search outcome. Techniques which improve the file have been considered in this paper and mathematical expressions have been derived which show that under quite general conditions one can improve search by associative adjustment of the file. It must be presumed that most file errors are errors of underindexing and that queries typically search for the presence of an indexed term rather than its absence.

* Research leading to this paper was sponsored by the Air Force Office of Scientific Research of the Office of Aerospace Research.

ASSOCIATIVE CORRECTION FOR UNDERINDEXING

1. INTRODUCTION

The term "association" with respect to document storage, search, and retrieval is subject to many interpretations--a fact which has not fostered the development of a rational theory of association. In a sense, all document search problems can be characterized as association problems. When one indexes he associates index terms with documents, when he classifies he associates "like" documents, and when he searches he associates documents with the presumed need of the user. However, the word association in the documentation field has come to imply a mathematical or statistical means by which the user is lead to consider a broader spectrum of documents than he would otherwise consider and which, at the same time, sharpens his attention to those documents of highest presumed interest to him. The necessary conditions for the accomplishment of these apparently opposing objectives have not been examined mathematically, and an initial attempt at one aspect of this analysis forms the principal subject matter of this paper.

Associative techniques have been considered for the accomplishment of two principal objectives: (1) to classify a collection into similar groups as an assistance to the searcher, and (2) to assign a "relevance number" to each document in a file, the relevance numbers presumably reflecting the searcher's interest in each document. Note that accomplishment of the first objective may be independent of specific searches, but the second is oriented to the user's needs on a given search. The two objectives are not independent, nor do they cover every possible application. They are the one point most discussed in the literature, however.

In the category of classification techniques are the clumping techniques of Needham and Jores [1] and Dale, Dale, and Pendergraft [2], as well as the classification techniques of Maron [3], Borko and Bernick [4], and Baker [5]. The work of Giuliano and Jones [6] perhaps best typifies the use of association to order the items in the file according to their presumed relevance to the request. In their formulation, the retrieval vector $\underline{r}$ is the product of three matrices and a query vector, $\underline{q}$, as follows:

$$\underline{r} = DCA\underline{q} \tag{1}$$

where, with a file of d documents and t terms, D is a d x d document-document connection matrix, C is a d x t document-term connection matrix (the indexed file), A is a t x t term-term connection matrix, and q is a t component query vector. D and A are linear transformations which account for associations among documents and terms, respectively. In this paper we consider the construction of the matrix A and a different process for matching the query $\underline{q}$ against the transformed file CA.

It is clear from (1) that the association matrix, A, may be considered as a right hand transformation of the index matrix C, or a left hand transformation of the search vector, $\underline{q}$. For a given transformation, the two views are indistinguishable. However, from the standpoint of estimating the transformation required to accomplish one's objective, there may be some importance to the distinction. There may even be some advantage in considering A to be the product of two transformation matrices (Bryant [7]), the first a right hand transformation of C and the second a left hand transformation of q.

There is a non-trivial difference in considering A to be a transformation of C or of q. If it is a transformation of C, the association technique leads one to documents in which the exact terms of the query were not indexed. If it is a transformation of q the association technique leads one to documents in which the indexed terms were not asked for. There are implied assumptions about the accuracy of indexing and searching in these two approaches. In this paper we concentrate on adjusting the index matrix C.

2. SOME PRELIMINARY CONSIDERATIONS

In formulation (1), above, the index matrix C is presumed to be a d x t matrix in which the entry c_{ij} represents the "relatedness" of the jth index tag to the ith document. The system permits the identification of parts of documents, rather than entire documents, but we will always refer to the rows of the matrix as "documents." Similarly, we will refer to the columns as "terms," although they may include words taken from the text, with or without indicated linguistic associations, and descriptors modified by the application of roles, links, or interfixes. The cell entry c_{ij} may, in general, be any real number, but frequently is either one or zero depending upon whether the jth term has or has not been selected by the indexing system. An indexing system which assigns values other than dichotomous values will be referred to as a "weighted indexing system" since, presumably, the variable quantity assigned is related to the strength of the relationship. Such a system perhaps is typified by one which is derived from word frequency counts.

In this paper we deal specifically with the index matrix of zeros and ones. In addition to the necessity to restrict the scope of the problem there are two cogent reasons for the choice of the unweighted index. First, it is very common, and large collections have been indexed in this way, e.g., the DDC collection. Second, there is an intuitive appeal to this kind of indexing on the grounds that the importance of the term to the document is derived from the needs of the searcher. The indexer cannot anticipate in advance when a concept which seems not to be relevant to the principal topic of the paper will be the exact thing the searcher is looking for. He may as well simply record its existence or its absence with a one or a zero.

The "relevance number," r_i, proposed by Giuliano and Jones is the dot product of two vectors, one of which has been transformed. One can easily show that this response measure need not be indicative of the closeness of the match between the two vectors unless certain conventions on scaling are adopted. We propose the use of a "measure of mismatch" defined as follows:

$$r_{ik} = \sum_j w_j (c_{ij} - q_{jk})^2 \qquad (2)$$

where r_{ik} is the measure of mismatch for the ith document with regard to the kth search, c_{ij} is the (possibly transformed) indexing of the ith document with respect to the jth term, q_{jk} is the kth search specification for the jth term, and w_j is a set of weights specified by the searcher (which, in some cases, will be all equal to one). If $w_j = 1$ for all j, then $\sqrt{r_{ij}}$ corresponds to the usual distance measure in n dimensional Euclidean space.

A further point requiring clarification is that the searcher may specify a subset of the total terms which, if indexed, indicate his interest in the document. He may also specify a set of terms which indicate his lack of interest. The former may appear in the search query as ones and the latter as zeros. He is presumed to be indifferent to the remainder of the possible terms, so they should not appear in the matching operation. In practice, one will only match a subset of the total terms, but, mathematically, one can handle this situation by assigning $w_j = 0$ for all "indifferent" terms.

It is well known that the assignment of index terms is subject to a great deal of error [8]. These errors can be characterized as errors of overindexing or underindexing [9], the latter being most common in practice [10]. The effects of errors of both kinds have been investigated, both theoretically [8] and empirically [10] for searches expressed as intersections of terms.

Two cases must be distinguished. There are situations in which an intelligent person, completely familiar with the subject matter and the indexing rules prescribed, would be able to say, with extremely small error, whether a given term should or should not be indexed for a given document. An example might be the indexing of an organic chemical compound, where the structure of the compound is given in the document. The other case encompasses a wider class of documents, such that among well informed indexers there would be substantial disagreement concerning the applicability of a given term to a specified document. While the consensus of a committee of experts might be defined as the correct indexing, such a procedure may be too artificial to be useful in practice. In this paper we consider the first case.

3. USE OF ASSOCIATION TO CORRECT UNDERINDEXING

Consider the indexing of the jth term to the ith document. A particular indexer will assign either a one or a zero, depending upon a number of factors. Comparison of the actual indexing with the "correct" indexing may yield the following categorization of responses, depending upon whether the given term should or should not have been assigned:

	Correct Indexing	
Actual Indexing	Don't assign term	Do assign term
Term not assigned	p_0^{ij}	p_1^{ij}
Term assigned	p_2^{ij}	p_3^{ij}
	1.0	1.0

Meaning can be ascribed to the symbols in the cells as follows: Suppose the jth term should not be assigned to the ith document. The relative frequency with which it would be assigned by a large population of indexers is represented by p_2^{ij}, while $p_0^{ij} = 1 - p_2^{ij}$. With some reasonable assumptions concerning the convergence of this relative frequency as the number of indexers increases it is appropriate to refer to p_2^{ij} as the "conditional probability of overindexing" with reference to the ith document and jth term. The condition, of course, is that the term should not have been indexed. Similarly, if the jth term should have been indexed, p_1^{ij} may be thought of as the "conditional probability of underindexing." Here, the condition is that the term should have been indexed.

It has been found useful to consider p_0^j, p_1^j, p_2^j, and p_3^j which have the meanings ascribed above to p_0^{ij}, p_1^{ij}, p_2^{ij}, and p_3^{ij}, except that the relative frequencies are averaged over a large collection of documents in the file.

Empirically, it has been observed that p_1^j tends to be substantially larger than p_2^j . This disproportion seems reasonable, since indexing is either a searching operation or a recognition process. In either case, errors of omission will tend to predominate.

In order to complete the characterization of the probability of underindexing we need to know something about the frequency with which the jth term <u>should be</u> indexed. Let y^j be the relative frequency with which an expert indexer would select the jth term, over all documents in the file. It is useful to identify this relative frequency with the prior probability that the jth term should be indexed.

With the definitions and conventions established above one can write for the ith document drawn at random from the file,

$$\begin{aligned} P(c_{ij} = 0) &= (1 - y^j)(1 - p_2^{ij}) + y^j p_1^{ij} \\ P(c_{ij} = 1) &= (1 - y^j)\, p_2^{ij} + y^j\,(1 - p_1^{ij}) \end{aligned} \qquad (3)$$

The c_{ij} are the entries of the index matrix C which we wish to adjust for association among terms. Denote by b_{ij} a value chosen to represent a revised measure of the relationship of the jth term to the ith document. We consider later how such an estimate might be derived, but, for the moment, we assume that such a figure can be obtained and, further, that it is unbiased. We will presume that we are interested only in correcting for underindexing errors. That is, if $c_{ij} = 1$ it will never be replaced by another figure, but if $c_{ij} = 0$ we may wish to replace it. Under these conditions, unbiasedness implies that

$$E(b_{ij} \mid c_{ij} = 0) = P(u_{ij} = 1 \mid c_{ij} = 0) \qquad (4)$$

where u_{ij} is the "correct" indexing, e.g., the indexing which the user might have chosen had he indexed the ith document with his specific search needs in mind.

By Bayes' Theorem

$$P(u_{ij} = 1 \mid c_{ij} = 0) = \frac{P(u_{ij}=1)P(c_{ij}=0 \mid u_{ij}=1)}{P(u_{ij}=0)P(c_{ij}=0 \mid u_{ij}=0)+P(u_{ij}=1)P(c_{ij}=0 \mid u_{ij}=1)}$$

$$= \frac{y^j\, p_1{}^j}{y^j\, p_1{}^j + (1 - y^j)(1 - p_2{}^j)} \qquad (5)$$

$$= v_j \text{ (for convenience in notation)}$$

where $p_1{}^j$, $p_2{}^j$, and v_j are to be interpreted as average values over the d documents in the file.

Let

$f_0(b_{ij})$ = the conditional density of b_{ij}, given $u_{ij} = 0$, with mean μ_{oj} and variance σ_{oj}^2

$f_1(b_{ij})$ = the conditional density of b_{ij}, given $u_{ij} = 1$, with mean μ_{1j} and variance σ_{1j}^2.

The effectiveness of the indexing system must be judged with regard to the queries which are put to it. However, since we are seeking to transform the indexing rather than the query we will assume that the query is error free in the sense that the searcher has complete knowledge of the indexing system (but not of the exact tags assigned) and has prepared his query in such a way as to maximize the disparity between measures of mismatch for unwanted documents. For definitional purposes the extent to which a document is "wanted" is determined by how nearly it matches the query. It is recognized that in actual cases queries are frequently poorly constructed and that searchers learn to accommodate their queries to the weaknesses of the indexing, but it is necessary to fix something in an otherwise totally fluid system. We have chosen to fix the searches.

By the above heuristic argument it seems reasonable to measure the effectiveness of the indexing of a particular term as the ratio between the expected contribution to mismatch for unwanted documents and the expected contribution to mismatch for wanted documents. Let ME_c be this measure of effectiveness for the unadjusted indexing, c_{ij}, and ME_b be the measure of effectiveness for the adjusted indexing, b_{ij}. Then the gain, G, due to adjustment can be expressed as

$$G = ME_b - ME_c \tag{6}$$

where

$$ME_b = \frac{E_{q_{jk}} E_{u_{ij}} \left[E(b_{ij} - q_{jk})^2 \,\middle|\, q_{jk} \neq u_{ij}\right]}{E_{q_{jk}} E_{j_{ij}} \left[E(b_{ij} - q_{jk})^2 \,\middle|\, q_{jk} = u_{ij}\right]} \tag{7}$$

$$ME_c = \frac{E_{q_{jb}} E_{u_{ij}} \left[E(c_{ij} - q_{jk})^2 \,\middle|\, q_{jk} \neq u_{ij}\right]}{E_{q_{jk}} E_{u_{ij}} \left[E(c_{ij} - q_{jk})^2 \,\middle|\, q_{jk} = u_{ij}\right]} \tag{8}$$

and

$E_{q_{jk}}$ = expectation over all values of q_{jk}, i.e.,

$P(q_{jk} = 1) = P_j$, and $P(q_{jk} = 0) = 1 - P_j$

$E_{u_{ij}}$ = expectation over all values of u_{ij}, i.e.,

$P(u_{ij} = 1) = V_j$ and $P(u_{ij} = 0) = 1 - V_j$

Some algebra shows that

$$ME_b = \frac{P_j((1-V_j)/V_j) \left[\sigma_{oj}{}^2 + (1-\mu_{oj})^2\right] + (1-P_j)(\sigma_{1j}{}^2 + \mu_{1j}{}^2)}{(1-P_j)(1-V_j)/V_j)(\sigma_{oj}{}^2 + \mu_{oj}{}^2) + P_j \left[\sigma_{1j}{}^2 + (1-\mu_{1j})^2\right]} \tag{9}$$

$$ME_c = (1 - V_j)/V_j \tag{10}$$

By definition, P_j is the relative frequency with which the <u>presence</u> of the jth term is sought rather than its <u>absence</u>. Thus, if the search queries contain only ones for the terms used in the search (rather than ones and zeros) $P_j = 1$ and (9), above, loses the second term of the numerator and the first term of the denominator. Also, by the unbiasedness conditions imposed on b_{ij},

$$V_j \mu_{1j} + (1 - V_j) \mu_{oj} = V_j \tag{11}$$

Using this relationship and imposing the further condition that $\sigma_{oj}^2 = \sigma_{1j}^2$, one finds that ME_b is greater than ME_c whenever μ_{1j} is greater than μ_{oj}. That is, one, under the above assumptions above P_j, σ_{oj}^2 and σ_{1j}^2, will expect to gain whenever $c_{ij} = 0$ can be replaced by a value b_{ij} which on the average is greater when the term should be indexed than when it shouldn't be indexed. These are unusually mild requirements and seem to dictate the extensive use of associative techniques to adjust for underindexing.

If $P_j < 1$, i.e., if one sometimes searches for documents which do not contain the jth term, the results are not so clear. They depend critically upon the values for P_j and V_j (the probability that the jth term should be indexed, given that it has not been indexed). Indifference curves, showing the values for μ_{1j} and σ_j^2 for which it is immaterial whether one adjusts or not are shown in Figs. 1 - 5 for values of V_j from 0.01 to 0.30 and values of P_j from 0.5 to 0.9. One will expect to gain if the values of μ_{1j} and σ_j^2 lie below and to the right of the displayed curves.

The quantities P_j, V_j, μ_{1j} and σ_j^2 are parameters which are unknown in given applications. However, they may be estimated through sampling.

Suppose one draws a random sample of n documents from the file and has them reindexed by experts. These may be the same documents used in estimation of the associative adjustments (Sec. 4). He then has the original indexing c_{ij} and the "correct" indexing u_{ij} . Consider the jo^{th} term and sort out the n_{jo} documents for which $c_{ijo} = 0$. The fraction of these for which $u_{ijo} = 1$ is an estimate of V_{jo} . Application of the estimating procedure (see Sec. 4) to the n_{jo} sample documents will yield n_{jo} estimates b_{ijo} from which μ_{ojo} , μ_{1jo} and σ_{jo}^2 can be estimated. P_j can be estimated from observations of recorded searches. In applying the sample estimates to Figs. 1 - 5 one may take into account by straightforward application of statistical methods the sampling variation in the estimation of the parameters.

4. ESTIMATION OF THE ADJUSTMENTS, b_{ij}

So far, nothing has been said about the way in which we estimate b_{ij} . The results of the previous section are independent of the manner of estimation.

It is helpful to consider the sources of information about the "correct" indexing and the nature of the errors associated with them. We have, first of all, the original indexing, c_{ij} , which we must presume provides information about the correct indexing. Second, it seems intuitive that u_{ij} (the correct indexing) is more likely to be 1 if the jth term is indexed with high frequency, or if the ith document has many terms indexed. Thus the marginal frequencies should provide information about the correct indexing. Finally, relationships among the indexings of the other terms should provide information about the indexing of the given term. We will refer to these three estimates as (1) the indexer estimate, (2) the marginal estimate, and (3) the regression estimate.

The indexer estimate is simply the indexing c_{ij} . Since we are only interested in correcting for underindexing we may adjust $c_{ij} = 0$ upward, but will never adjust $c_{ij} = 1$ downward. Therefore, the indexer estimate, for the cases we wish to adjust, is always 0. Thus, its variance is also zero. However, its mean square error (which takes into account bias as well as variability) can be estimated for the jth term by

$$S^2_{1j} = \frac{\sum_i (c_{ij} - u_{ij})^2}{n_{oj}} = n_{uj}/n_{oj} \tag{12}$$

where n_{oj} = the number of zero indexings of the jth term in a sample of the file and n_{uj} = the number of corrected zero indexings in the same sample.

The marginal estimate for the ijth entry of the index matrix can be found as follows:

$$m_{ij} = \frac{(\sum_j c_{ij})(\sum_i c_{ij})}{\sum_{ij} c_{ij}} \tag{13}$$

If this estimate has any predictive power, then

$$\overline{m}^{(1)}_{ij} - \overline{m}^{(0)}_{ij} > 0 \tag{14}$$

where $\overline{m}^{(1)}_{ij}$ is the average value of the marginal estimates for the cells for which $c_{ij} = 1$ and $\overline{m}^{(0)}_{ij}$ is the average value for the cells for which $c_{ij} = 0$. One can take as the marginal estimate, for predictive purposes

$$m^{(0)}_{ij} = c\, m_{ij} \tag{15}$$

where c is so chosen as to make the average value of $m^{(0)}_{ij}$ equal to v_j (to be determined empirically from the sample file). An estimate of the variance of the marginal estimate is then

$$S^2_{2j} = \frac{\sum\limits_i (m^{(0)}_{ij} - u_{ij})^2}{n_{oj}} \tag{16}$$

An alternative marginal estimate is given by

$$m'_{ij} = \frac{\sum\limits_j c_{ij}}{n_1} + \frac{\sum\limits_i c_{ij}}{n_j} - \frac{\sum\limits_{ij} c_{ij}}{n} \tag{17}$$

This formulation assumes that the indexings are an additive function of the marginal means, whereas expression (13) assumes that indexings are proportional to marginal means. Not enough information has been gathered empirically to judge which estimate is better.

The regression estimate is determined for each term by applying standard linear regression techniques to the indexings of a subset of the other terms in the document. It is presumed that expert judgment can be called on to select terms which may have predictive power, as well as to provide Boolean functions of the indexings of various combinations of terms to be tested in the regression studies. The resulting estimate has its own built-in estimate of the variance, which we will denote by S_{3j}^2 .

Denote the indexer estimate by x_{1ij}, the marginal estimate by x_{2ij} , and the regression estimate by x_{3ij} . Then

$$b_{ij} = \lambda_1 x_{1ij} + \lambda_2 x_{2ij} + \lambda_3 x_{3ij} \tag{18}$$

where the λ_k are weights, so chosen that they sum to 1 and so that they minimize the variance of b_{ij} . If the estimates x_{1ij}, x_{2ij}, and x_{3ij} are independent, then

$$\lambda_1 = \frac{S_{2j}^2 S_{3j}^2}{S} \tag{19}$$

$$\lambda_2 = \frac{S_{1j}^2 S_{3j}^2}{S}$$

$$\lambda_3 = \frac{S_{1j}^2 S_{2j}^2}{S}$$

where $S = S_{1j}^2 S_{2j}^2 + S_{1j}^2 S_{3j}^2 + S_{2j}^2 S_{3j}^2$. In case the estimates are correlated (which can be determined from a sample of documents) a different weighting technique must be applied.

For convenience in notation we drop the subscripts $\underline{i}$, $\underline{j}$ and write

$$b = \lambda_1 x_1 + \lambda_2 x_2 + \lambda_3 x_3 \tag{20}$$

Then,

$$\begin{aligned} \text{Var } b = {} & \lambda_1^2 \text{Var } x_1 + \lambda_2^2 \text{Var } x_2 + \lambda_3^2 \text{Var } x_3 + 2\lambda_1\lambda_2 \text{Cov } x_1 x_2 \\ & + 2\lambda_1\lambda_3 \text{Cov } x_1 x_3 + 2\lambda_2\lambda_3 \text{Cov } x_2 x_3 \end{aligned} \tag{21}$$

One minimizes Var b + $2\gamma(\lambda_1 + \lambda_2 + \lambda_3 - 1)$ with respect to λ_1, λ_2, λ_3, and γ to achieve the following equations in matrix form:

$$\begin{bmatrix} \text{Var } x_1 & \text{Cov } x_1x_2 & \text{Cov } x_1x_3 & 1 \\ \text{Cov } x_1x_2 & \text{Var } x_2 & \text{Cov } x_2x_3 & 1 \\ \text{Cov } x_1x_3 & \text{Cov } x_2x_3 & \text{Var } x_3 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \gamma \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \qquad (22)$$

One can insert sample values for the variances and covariances and solve by any of the usual methods to obtain the approximate weights $\hat{\lambda}_1$, $\hat{\lambda}_2$, and $\hat{\lambda}_3$ to be used in the process of averaging the three separate estimates. Let V represent the variance-covariance matrix, above and V^{-1} its inverse. Then, it may be seen that $\lambda_1 = V^{14}$, $\lambda_2 = V^{24}$, and $\lambda_3 = V^{34}$ where V^{14} indicates the first-row, fourth-column element of the inverse, and so on.

A point worth noting is that associative adjustment of the file can be accomplished by the information system's computing center during slack time. Further, it need not be done all at one time, but can be done piecewise, either by adjusting a few terms at a time or a larger collection of terms for a subset of documents. There is no implication that the adjustment need be made to all terms in the file in order to be effective. It is clear from Figs. 1 - 5 that one can gain most dramatically by adjusting first those for which there is high underindexing error (i. e., V_j is high) and for which predictability is high (i. e., μ_{1j} is high and σ_j^2 is low). There is surely a point beyond which the cost of associative adjustment would exceed possible gains to be derived.

Some extension of associative adjustment is possible to cases where it is not feasible to define a "correct" indexing [11] . In these cases it appears to have considerable merit as well.

SELECTED REFERENCES

1. R. M. Needham and K. Sparck Jones, "Keywords and Clumps," The Journal of Documentation, Vol. 20, No. 1, March 1964, pp. 5-15.

2. A. G. Dale, N. Dale and E. D. Pendergraft, "A programming system for automatic classification with applications in linguistics and information retrieval research," prepared for NSF GN-208, Linguistics Research Center, The University of Texas, Austin, October 1964.

3. M. E. Maron, "Automatic indexing: An experimental inquiry," JACM 8, 407-417, 1961.

4. Harold Borko and Myrna Bernick, "Automatic document classification," JACM 10, 151-162, 1963.

5. Frank B. Baker, "Information retrieval based upon latent class analysis," JACM 9, 512-521, 1962.

6. Vincent E. Giuliano and Paul E. Jones, "Linear associative information retrieval," Chapter 2 of Howerton and Weeks, Vistas in Information Handling, Vol. 1, Spartan Books, Washington, D. C., 1963.

7. Edward C. Bryant, "Some notes on associative retrieval," U. S. Department of Commerce, Patent Office, January 1964, PB 166 511.

8. J. Jacoby and V. Slamecka, "Indexer consistency under minimal conditions," report to Rome Air Development Center, Contract AF 30(6D2)-2616, Documentation, Inc., Bethesda, Nov. 1962.

9. Edward C. Bryant, Donald W. King, and P. James Terragno, "Some technical notes on coding errors," WRA PO 7, report to the U. S. Dept. of Commerce, Patent Office, Contract Cc6078, Westat Research Analysts, Inc., Denver, July 1963, PB 166 487.

10. Edward C. Bryant, Donald W. King, and P. James Terragno, "Analysis of an indexing and retrieval experiment for the organometallics file of the U. S. Patent Office," WRA PO 10, Report to U. S. Dept. of Commerce, Patent Office, Contract Cc 6078, Westat Research Analysts, Inc., Denver, August 1963, PB 166 488.

11. Edward C. Bryant, Donald T. Searls, and Robert H. Shumway, "Some theoretical aspects of the improvement of document screening by associative transformations," report to AFOSR under contract AF 49(638)-1484, Westat Research Analysts, Inc., Denver, November 30, 1965 (AD 628 191).

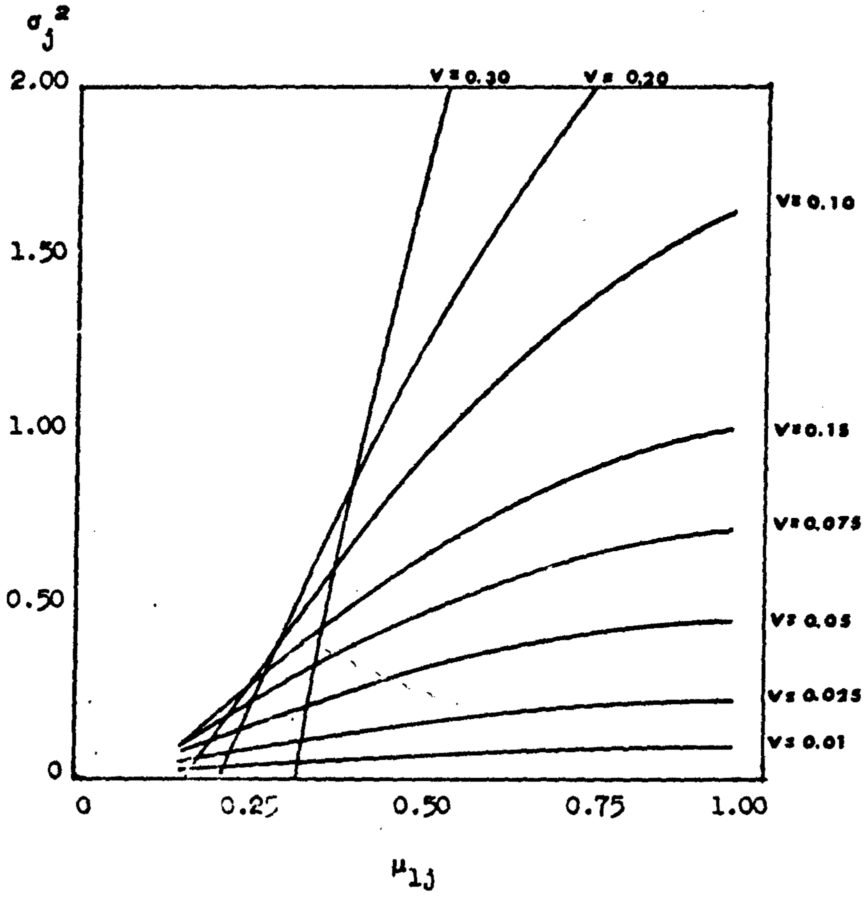

Fig. 1. Indifference Curves, P = 0.9

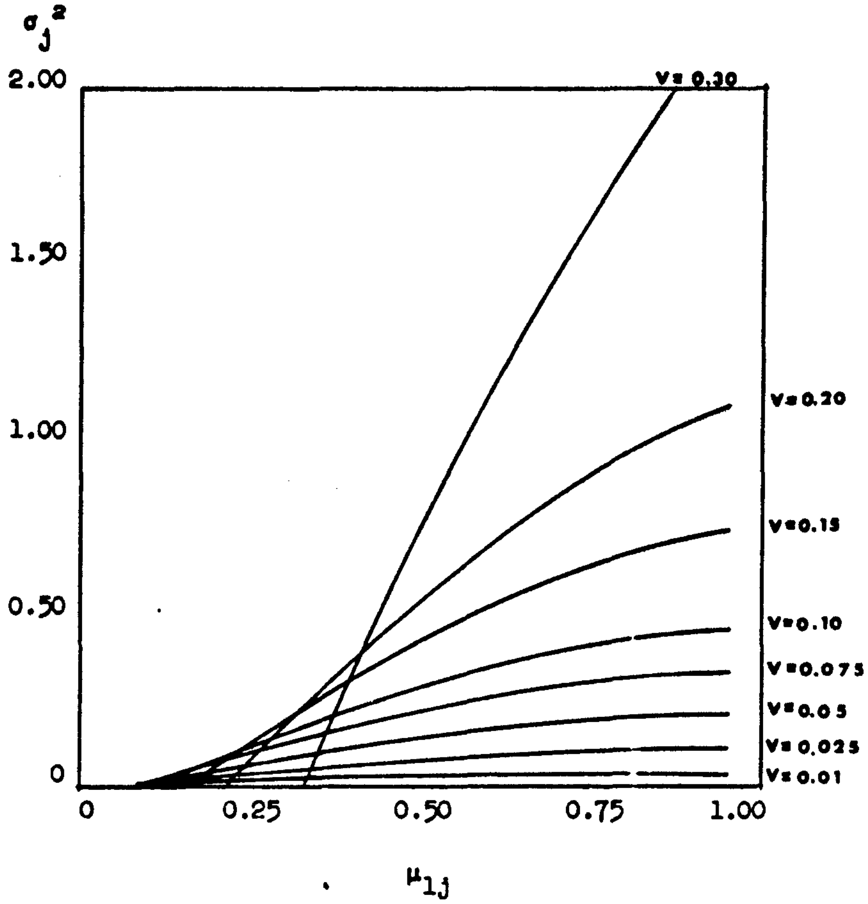

Fig. 2. Indifference Curves, P = 0.8

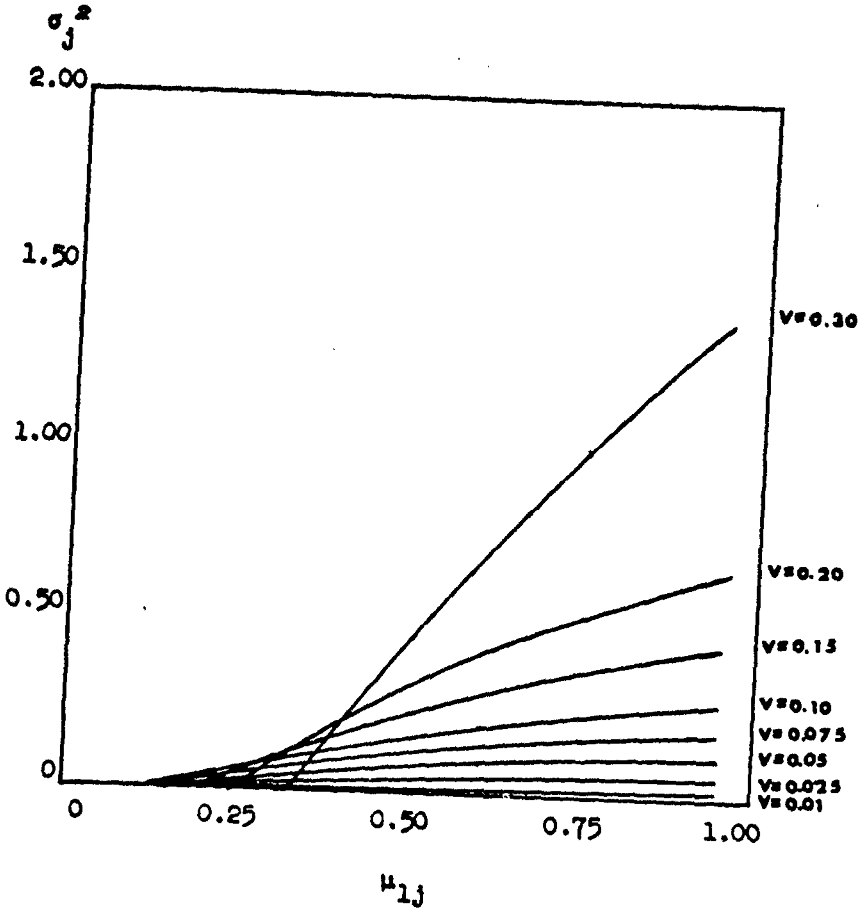

Fig. 3. Indifference Curves, P = 0.7

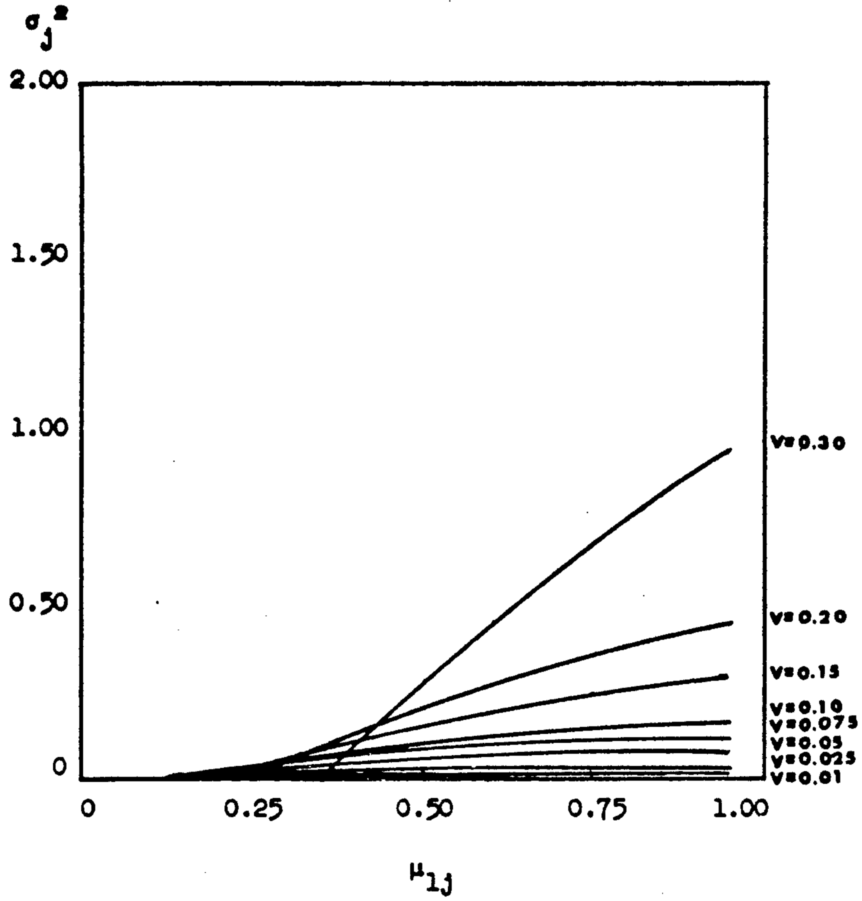

Fig. 4. Indifference Curves, P = 0.6

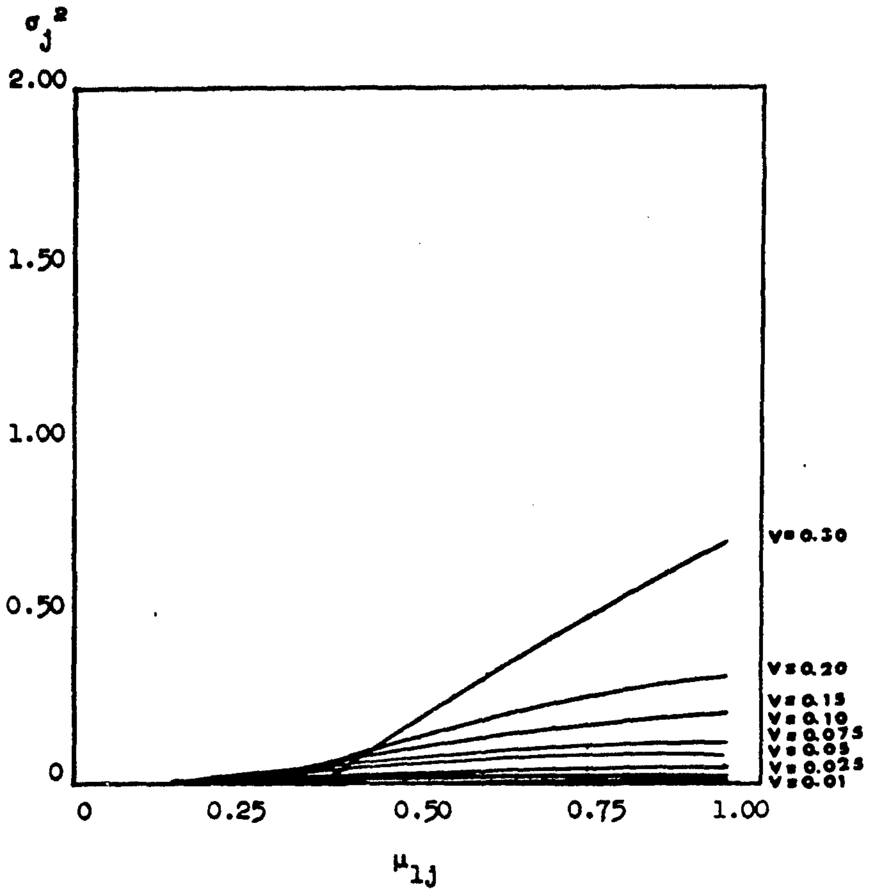

Fig. 5. Indifference Curves, P = 0.5

APPENDIX 2

ON THE EXPECTED GAIN FROM ADJUSTING MATCHED TERM RETRIEVAL SYSTEMS *

by

R. H. Shumway
Westat Research, Inc.

Abstract - A file adjustment procedure based on maximizing the Bayes expected gain is proposed for matched term retrieval systems. The expected gain and its probability distribution are derived as a function of 1) the prior proportion of omitted terms and 2) the coefficient of separation between two distributions corresponding to values of an adjustment statistic. An example evaluates the gain parameters for a typical information retrieval system.

INTRODUCTION

A number of papers (1) - (5) have been directed towards the problem of developing transformations or adjustments to be applied to term adjusted files. Generally the term associations are used to generate a set of adjusted codings which improve retrieval by leading one more quickly to the relevant documents. However, while many empirical evaluations have been made based on file adjustments made on experimental data, theoretical investigations into the amount that one could reasonably expect to gain in retrieval effectiveness from such procedures have been notably lacking. (An exception is reference (7)).

* Sponsored in part by the Air Force Office of Scientific Research of the Office of Aerospace Research and partly by the U.S. Department of Commerce, Patent Office.

It is the intention here to provide a possible basis within a decision theoretic framework for evaluating the gain which might be expected for some file adjustment procedures. The basic approach, as in (7), is to consider only adjustments which correct for term omissions using the empirical result that the relative frequency of incorrectly applied indexed terms is negligible (6). With this restriction we may limit our attention to developing an approach for deciding whether or not a term should be adjusted upward. This binary decision can be formulated in Bayesian terms with the probability of a user adjusting a term upwards playing the part of a prior probability. We use a measure which associates with each document (term) a measure of its association (mismatch) with the query. Our definition of gain is the amount that the measure of mismatch can be increased for irrelevant documents or decreased for relevant documents by making a set of corrections for underindexing. A procedure for adjustment is chosen which is optimal in the sense that it maximizes the gain and this gain is tabulated for various values of the system parameters. Finally we compute the probability distribution of the gain along with the positive gain probability. Thus, for binary adjustment procedures which assign either a zero or one to the corrected indexing we may evaluate the gain for systems in which the basic parameters can be measured.

THEORETICAL CONSIDERATIONS

We shall use the formulation of Bryant et al (6) as a basis for the theoretical development. In this case the original term indexed file is regarded as a dxt matrix of zeros and ones, say c_{ij}, with c_{ij} taking the value 1 if the jth term pertains to the ith document and 0 otherwise. We consider a set of requests or queries expressible as a matrix q_{jk} where q_{jk} is assigned a value of 1 if the searcher regards the presence of the jth term as important in the kth query and 0 otherwise. Hence, a measure of mismatch between the mth document and the kth query can be defined as:

$$r_{mk} = \sum_{j} (c_{mj} - q_{jk})^2 \qquad (1$$

If the c's and the q's are either 0 or 1, equation (1) reduces to the number of mismatched terms between the mth document and the kth query. This measure of mismatch gives one the option of asking for the absence of certain terms as well as their presence. Note that in equation (1) the summation is, in general, performed over a subset of terms which are of interest to the searcher.

We suppose now that the original indexings c_{ij} are not indexed correctly or at least they are not indexed from the point of view of the searcher or ideal user who might prefer to have assigned some different coding u_{ij}. We assume, as in (6), that underindexing represents the major type of error in the file and adjust only terms originally indexed with a 0. Let u_{ij} (0 or 1) be the value that the ideal user would assign. Suppose that it is not feasible to correct all the term indexings c_{ij} with the ideal user and that the correction is to be made on the basis of some statistic $T = T(c_{11}, c_{12}, \ldots, c_{dt})$ computed from the other unadjusted codings. We do not consider the method (associative or otherwise) for generating this statistic but regard it as being characterized by the two conditional probability distributions:

$$F_0(x) = P(T \leq x \mid u_{ij} = 0) \qquad F_1(x) = P(T \leq x \mid u_{ij} = 1) \qquad (2)$$

The first distribution function F_0 gives the probability distribution for the statistic T when the adjusted term should be 0 while F_1 gives the distribution of the statistic when the adjusted term should be 1. Figure 1 shows the possible forms which the density functions f_0 and f_1 corresponding to the distributions given in (2) could take. Our procedure for assigning a user indexing will be a binary decision scheme which assigns $u_{ij} = 1$ for $T > K$ and assigns $u_{ij} = 0$ for $T \leq K$ since we shal. presume that the statistic T chosen should be high when $u_{ij} = 1$ and low when $u_{ij} = 0$. The assigned user value will not always be identical to the correct user indexing so that to avoid confusion we will denote this assigned user indexing by b_{ij}.

Equation (1) indicates that the measure of mismatch is also influenced by the query indexing through the parameter q_{ij} which may take the values 0 or 1. Hence, the identities and values of a number of parameters associated with a single term may be arranged as in Table I. (In subsequent discussion of single term values the subscript ij is omitted.) The library coding c is always 0 since errors of overindexing are being neglected. In order to proceed further with the analysis of Table I, some assumptions are needed about the joint probability distributions of b, c, u and q and we assume that the user indexing u and the query are independent of each other and that the query q is independent of the adjusted coding b. Hence, the expected gain for a single

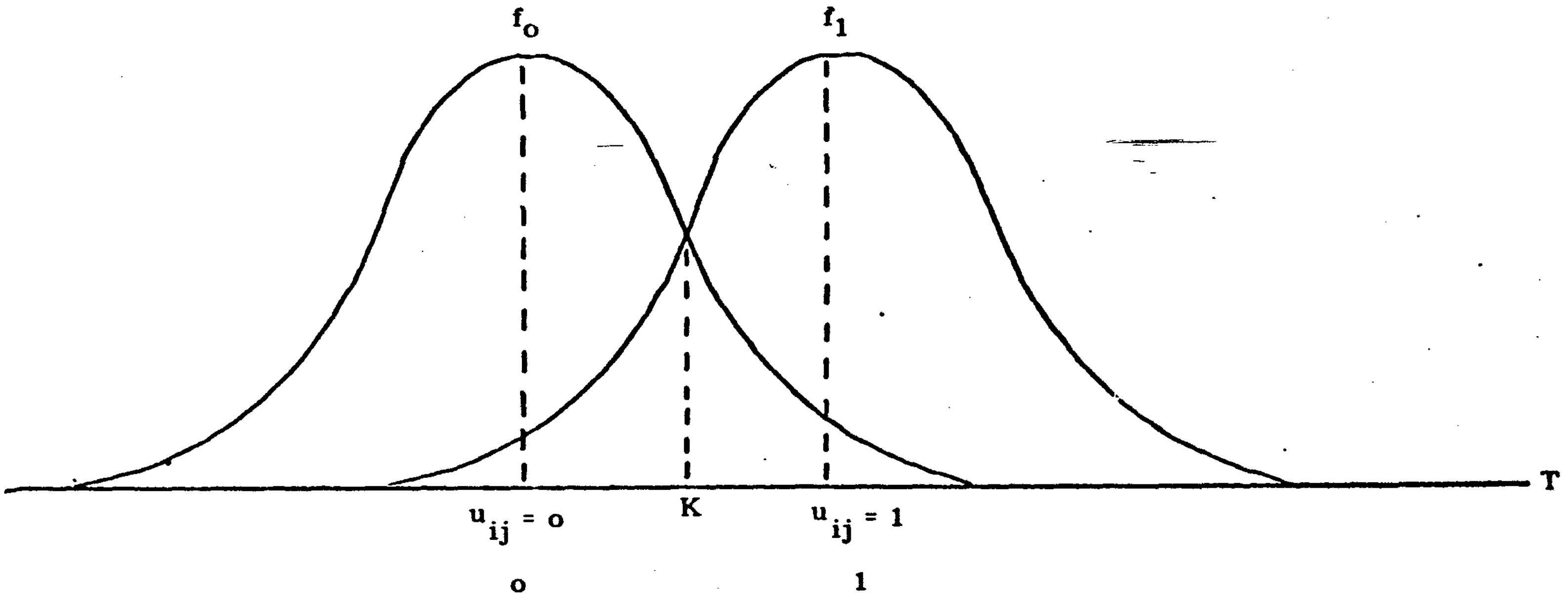

Figure 1 - Distribution of the statistic T for $u_{ij} = 0, 1$ respectively with $\sigma = .5$

TABLE I

System Parameters for File Adjustment

				(1)	(2)	(3)	
q	c	u	b	$(u - q)^2$	$(c - q)^2$	$(b - q)^2$	GAIN
0	0	0	1	0	0	1	-1
0	0	0	0	0	0	0	0
0	0	1	1	1	0	1	1
0	0	1	0	1	0	0	0
1	0	0	1	1	1	0	-1
1	0	0	0	1	1	1	0
1	0	1	1	0	1	0	1
1	0	1	0	0	1	1	0

(1) Desired contribution to mismatch

(2) Contribution to mismatch without adjustment

(3) Contribution to mismatch with adjustment

q Query indexing

u User indexing

b Adjusted indexing

c Original indexing

term search is expressed as:

$$E\ (G) = \sum_{b,u,q} G(b,u,q)\ P(b \mid u)\ P(u)\ P(q) \qquad (3)$$

where G(b, u, q) is some appropriate gain function defined for each b, u and q. The conditional distribution of b given u is determined by the decision point K in Figure 1 for:

$$\begin{aligned} P(b=0 \mid u=0) &= P(T \leq K \mid u=0) = F_0(K) \\ P(b=1 \mid u=0) &= P(T > K \mid u=0) = 1 - F_0(K) \\ P(b=0 \mid u=1) &= P(T \leq K \mid u=0) = F_1(K) \\ P(b=1 \mid u=1) &= P(T > K \mid u=1) = 1 - F_1(K) \end{aligned} \qquad (4)$$

We also take the densities of u and q to be given as binomial with parameters v and Q respectively. If the values of the parameters are examined, it is clear that the measure of mismatch and hence the ranking is influenced in a predictable way by the adjustment procedure. Our values of the gains filled in from columns (1), (2) and (3) of Table I reflect these considerations. For example, in the first row the desired contribution to mismatch $(u - q)^2$ is 0 with the contribution to mismatch without adjustment $(c - q)^2$ also being 0. The adjusted mismatch is 1 which is in error, contributing a gain of -1. The reader may easily convince himself that the other gains are reasonable and that positive gains tend to reflect a favorable adjustment of the mismatch and hence, the ranking. Then, using Table 1 and equations (3) and (4) with the binomial assumption on u and q leads to:

$$E(G) = -(1 - Q)(1 - v)(1 - F_0) + (1 - Q)v(1 - F_1) - Q(1 - v)(1 - F_0)$$
$$+ Q\,v\,(1 - F_1) = v\,(1 - F_1) - (1 - v)(1 - F_0) \qquad (5)$$

which is maximized by choosing a value K such that:

$$\frac{f_1(K)}{f_0(K)} = \frac{1 - v}{v} \qquad (6)$$

If the probability densities f_0 and f_1 are known or a discrete approximation is available, we may solve for K using equation (6) and then substitute into equation (5) to determine the maximum expected gain. For example, if the densities f_0 and f_1 can be regarded as being approximately normal with means 0 and 1 respectively and common variance σ^2, equation (6) yields:

$$K = 1/2 + \sigma^2 \log \frac{1 - v}{v} \qquad (7)$$

with the maximum expected gain per term represented in (5) as a function of σ^2 and v. In this case the mean separation is unity so that the value of σ represents a "coefficient of discrimination" in the sense that a larger σ is associated with an increased difficulty in discriminating between u = 0 and u = 1.

The above results pertain to single term searches only and it would be useful to extend the results to a search involving N terms. In addition, we are interested not only in the expected maximum gain but also in the exact or approximate probability distribution of the gains. The gain density for a single term search can be written down immediately from Table I and is reproduced below.

TABLE II

Probability Distribution of Gain for a Single Term Search

GAIN G	Probability Distribution P_G
-1	$(1 - v)(1 - F_0)$
0	$F_0 + v(F_1 - F_0)$
1	$v(1 - F_1)$

In an N term search the gain can range over the integers $-N$, $-N+1$, ..., 0, 1, ..., N. Then, let n_G be the number of terms in the search that produced a single term gain of G. Then, if the total gain is designated by G_T we may write

$$P(G_T = k) = \sum_{\substack{n_1 - n_{-1} = k \\ n_{-1} + n_0 + n_1 = N}} P_{-1}^{\,n_{-1}} P_0^{\,n_0} P_1^{\,n_1} \qquad (8$$

For moderate sized N, G_T will be the sum of the individual single term gains and the central limit theorem will apply yielding:

$$P(G_T \leq k) \cong \phi\left(\frac{k - \mu_T}{\sigma_T}\right) ; \quad \mu_T = N\, E(G), \quad \sigma_T = \sigma_G (N)^{1/2} \qquad (9)$$

an an approximate expression for the probability distribution of the gain.

Here $\phi(x)$ denotes the cumulative normal distribution with E(G) and σ_G the mean and standard deviation of the gain as computed from Table II. One measure of possible interest would be $P(G_T > 0)$ or the probability of making a positive gain. We shall henceforth refer to this measure as the "positive gain probability."

EXAMPLES

The measures of effectiveness developed in the preceding section will be quite different for the various adjustment procedures in both the form and separation of the distributions f_0 and f_1 of Figure 1. Empirical data categorizing adjusted and unadjusted terms into correctly adjusted and unadjusted terms and incorrectly adjusted and unadjusted terms, as well as the sample values T of the adjustment statistic will be needed in order to determine the performance characteristics of a particular system. Since the distribution of T is often the distribution of some linear combination of adjacent terms as in adjustment procedures using regression or other associative correction measures, we may frequently assume that it is approximately normal for terms that should have been adjusted as well as for terms that should not have been adjusted. For purposes of simplified computation we shall also assume in this example that the variances are equal in the two populations and that the average separation between f_0 and f_1 has been normalized to one. This allows the use of equation (7) to determine a cutoff point which maximizes the expected gain. Equation (5) then determines the maximum expected gain as a function of the parameters v and σ^2. Figure 2 shows the expected gain per term in the mismatch measure as a function of the prior proportion of omitted terms v and the spread of f_0 and f_1 denoted by σ^2. Note that we can never gain more on the average than the value of the parameter v. Also, with increasing σ the maximum expected gain goes down while with increasing v the maximum expected gain increases. If the basic parameters remain relatively constant from term to term the expected total gain from an N term search is N times the expected single term gain. Note that this expected gain is over terms in a single document which were not indexed in the original file. Hence, in a 20 term search a single document might contain only ten candidates for adjustment. Therefore, using Figure 2 with $v = .22$ and $\sigma = .5$ a maximum expected gain of $.10(10) = 1$ would be reasonable for documents containing ten terms originally indexed as zero.

In some cases a more interesting and informative measure might be the probability of making a specified gain, determined from equation (8) or its approximation (9). The characteristics of the system will determine the particular probabilities which contribute the most as measures of effectiveness. We have chosen to present the probability of making a positive gain $P(G_T > 0)$ in Figures 3 and 4. Note that while the expected gain increases with v and decreases with σ the probability of some gain (positive gain probability) for values of v less than .20 is increased with an increased variance. Hence, in this example, the improvement in expected gain with the decreased σ leads to a slight decrease in the positive gain probability. The phenomenon observed above where the expected gain and positive gain probability seem to

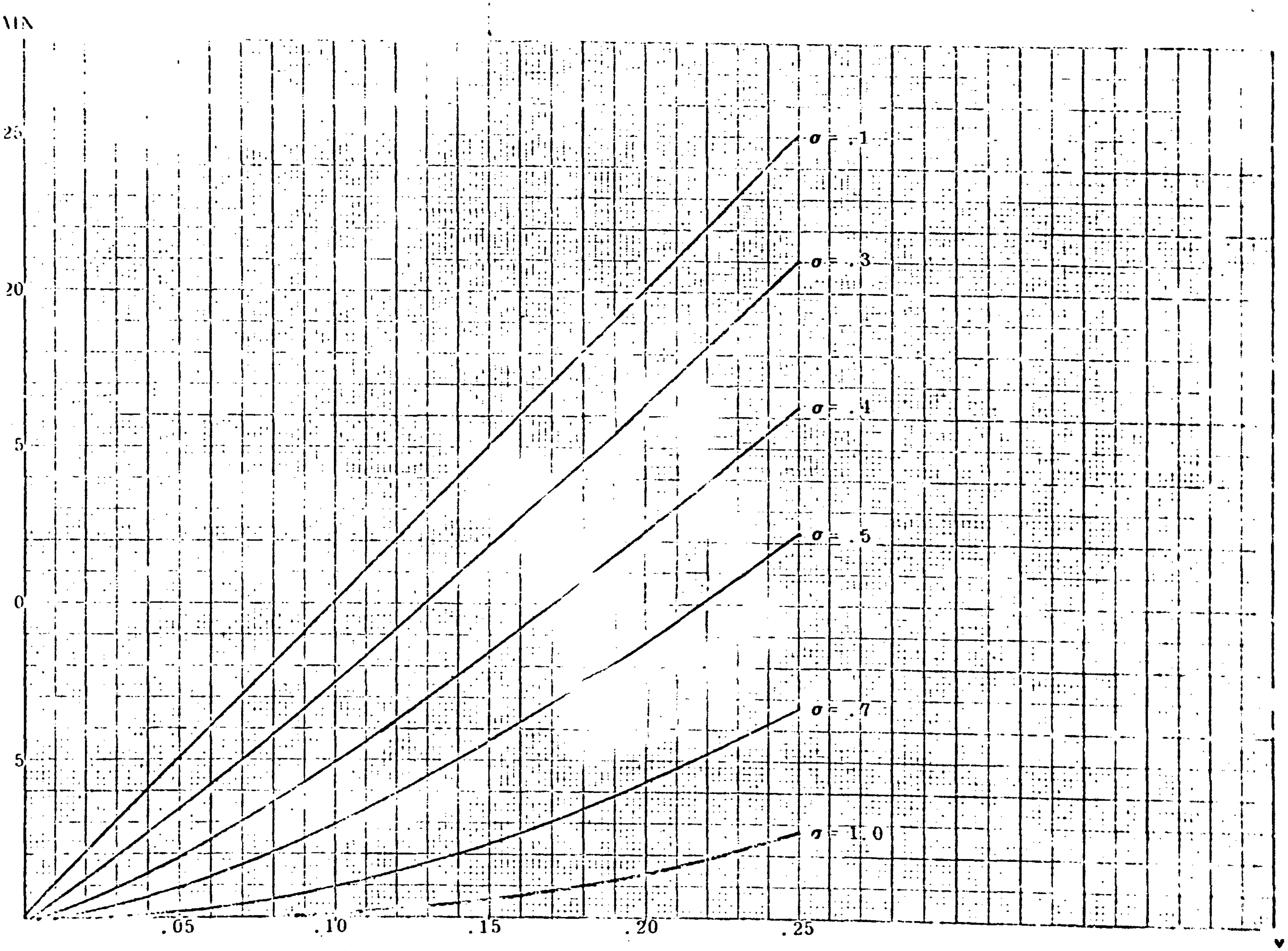

Figure 2 - Maximum expected gain as a function of v, the prior probability of adjustment.

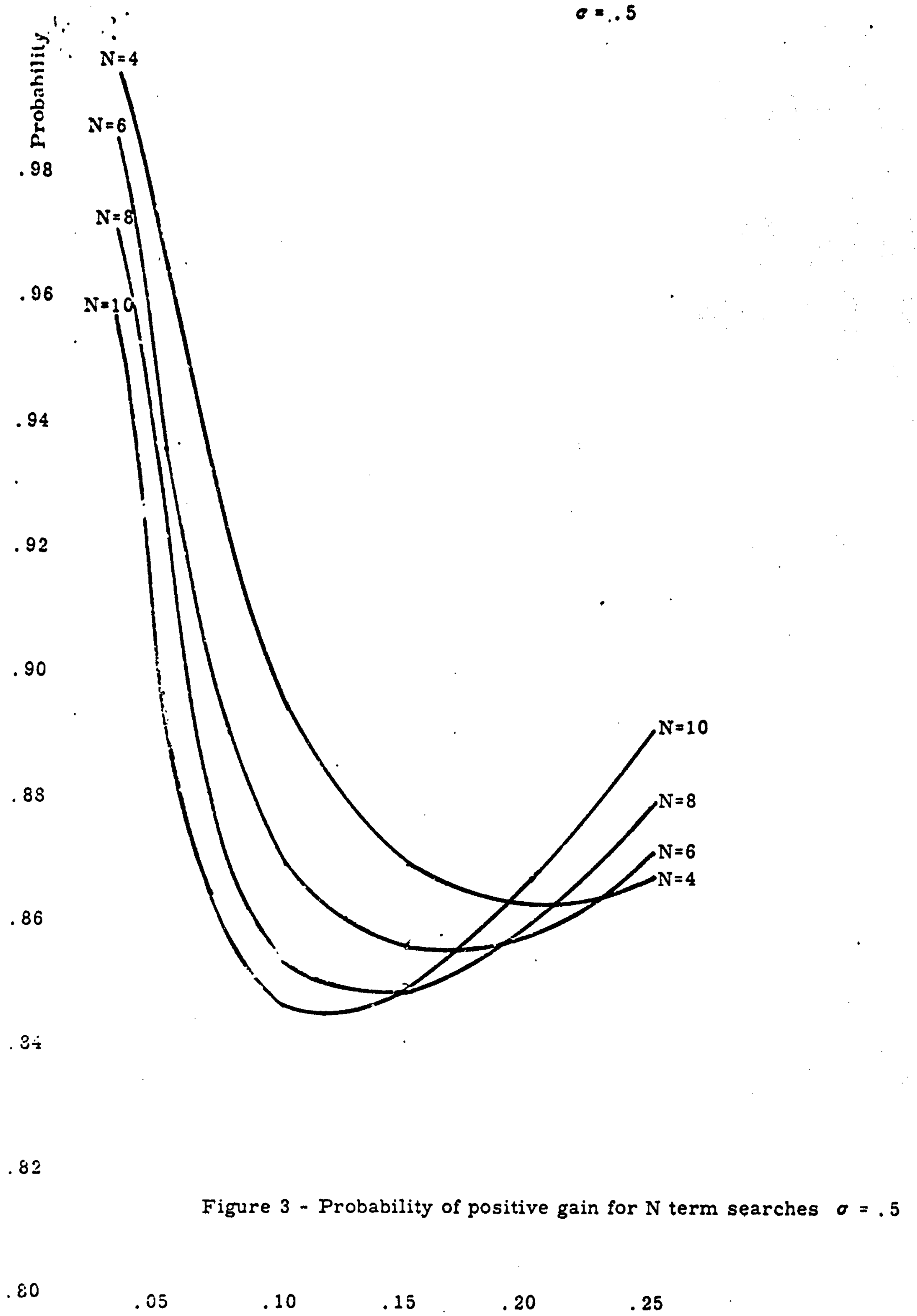

Figure 3 - Probability of positive gain for N term searches σ = .5

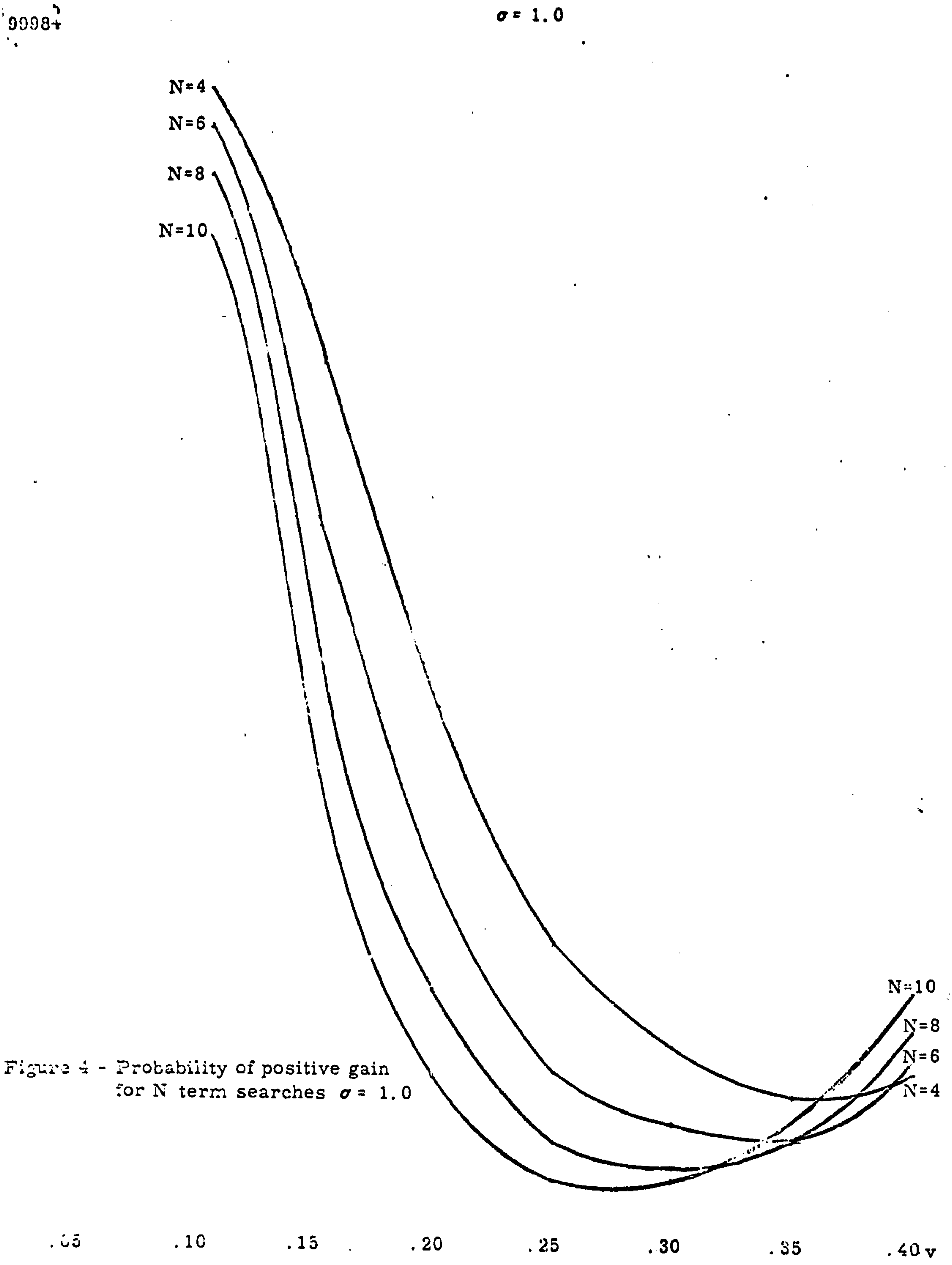

Figure 4 - Probability of positive gain for N term searches $\sigma = 1.0$

work against each other does not cause serious problems since the positive gain probability is uniformly high over the entire range of v. The same conflict characterizes the relation between the gain and the number of terms in the search with expected gain increasing for higher N and the positive gain probability decreasing. If the mean separation between the distributions in Figure 1 is positive we will always have a positive expected gain regardless of the variance σ^2.

As another example consider the computation of the entire probability distribution of the gain as given by equation (8). Let us suppose that in making five-term searches it is true on the average that three terms in the documents would not be coded in the unadjusted file. Assume also that the prior probability of omission, v, is .10. Then, for $\sigma = .5$ we use equation (8) to determine that the probability of gaining one is about .12. If we are searching for presence in the query then there is a .12 chance of decreasing the mismatch by one, which with a total possible mismatch of five would lead to a substantial improvement in the ranking. If the prior probability of omission is .20 the chance of a gain of one increases to .26. In this case the expected gain and gain probability do not seem to work against one another. It is also clear that the gain probability is a measure of the improvement in the ranking if it is assumed that a documents position in the ranking is determined incorrectly because of omitted terms.

CONCLUSIONS

We have developed the expected Bayes gain and the positive gain probability as measures of retrieval effectiveness for file adjustment procedures. These measures do not depend on the form of the adjustment which has been applied as it may be any one of a number of the so called associative schemes. The requirements are that the proposed procedure generate a set of adjustment statistics on a continuous scale and that the correct codings corresponding to these adjustments be available. Then the competing forms of Figure 1 can be plotted and the distributional forms F_0 and F_1 can be estimated. This yields a critical value K which maximizes the expected Bayes gain. The resulting measures of retrieval effectiveness (here the expected gain and the positive gain probability) are expressed in terms of the prior probability that a user would have preferred a different indexing. The computed examples show that it would be useful to examine the parameters in an operating system quite closely to determine the relative benefits of competing adjustment procedures.

REFERENCES

1. Baker, Frank B., Information Retrieval Based on Latent Class Analysis, JACM 9, 512-521, 1961.

2. Borko, Harold and Myrna Bernick, Automatic Document Classification, JACM 10, 151-162, 1963.

3. Giuliano, Vincent E. and Paul E. Jones, Linear Associative Information Retrieval, Chapter 2 of Howerton and Weeks, Vistas in Information Handling, Vol. 1, Spartan Books, Washington, D. C., 1963.

4. Maron, M. E., Automatic Indexing: An Experimental Inquiry, JACM 8, 407-417, 1961.

5. Needham, B. M. and K. Sparck Jones, Keywords and Clumps, The Journal of Documentation, Vol. 20, No. 1, March, 1964, 5-15.

6. Bryant, E. C., D. W. King and P. J. Terragno, Analysis of an Indexing and Retrieval Experiment for the Organometallics File of the U. S. Patent Office, P. O. 10, Report to U. S. Dept. of Commerce, Patent Office Contr. 6078, Westat Research, Inc. 1963, PB 166 488.

7. Bryant, E. C., D. T. Searls and R. H. Shumway, Associative Correction for Underindexing (submitted) (1966), Air Force Office of Scientific Research, Contr AF49(638)-1484, Nov. 30, 1965, AD 628 191.

APPENDIX 3

DERIVATION OF THE EXPECTATION AND VARIANCE OF THE GENERALIZED MISMATCH M

Let the generalized mismatch be defined by

$$M = (\underline{c} - \underline{q})' W(\underline{c} - \underline{q}) \qquad (1)$$

where $\underline{c}' = (c_1, c_2, \ldots, c_t)$ and $\underline{q}' = (q_1, \ldots, q_t)$ with $\underline{c}'$ a random vector of codings and $\underline{q}'$ a fixed query vector such that $E_R(\underline{c}) = \underline{q}$ for relevant documents and $E_I(\underline{c}) = \underline{q} + \underline{\epsilon}$ for irrelevant documents. W is a symmetric txt weighting matrix. Consider the derivation of the quantities $E_R(M)$, $E_I(M)$, $\mathrm{var}_R(M)$ and $\mathrm{var}_I(M)$ which are the means and variances of the mismatch for relevant and irrelevant documents. We may immediately write

$$E_R(M) = \mathrm{tr}\, W\Sigma \qquad (2)$$

$$E_I(M) = E_I(\underline{c} - \underline{q} - \underline{\epsilon})' W(\underline{c} - \underline{q} - \underline{\epsilon}) = \mathrm{tr}W\Sigma + \underline{\epsilon}' W\underline{\epsilon} \qquad (3)$$

where

$$\Sigma = \left\{\sigma_{ij}\right\} = E_R(\underline{c} - \underline{q})(\underline{c} - \underline{q})' = E_I(\underline{c} - \underline{q} - \underline{\epsilon})(\underline{c} - \underline{q} - \underline{\epsilon})' \qquad (4)$$

To develop the variance it is assumed that the $\underline{c}$ is a vector of jointly normally distributed variates so that

$$E_R(M^2) = \sum_{ijkl} E_R(c_i - q_i)(c_j - q_j)(c_k - q_k)(c_l - q_l)W_{ij}W_{kl}$$

$$= \sum_{ijkl} (\sigma_{ij}\sigma_{kl} + \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk})W_{ij}W_{kl} \qquad (5)$$

using the fourth moment law for normal random variables (2). Since W is symmetric $W_{kl} = W_{lk}$ in the second term and remembering that $\sigma_{ij} = \sigma_{ji}$ the above may be written

$$E_R(M^2) = (\mathrm{tr} W\Sigma)^2 + 2\mathrm{tr}(W\Sigma)^2$$

or

$$\mathrm{var}_R(M) = 2\mathrm{tr}(W\Sigma)^2$$

Now

$$E_I \left\{(\underline{c} - \underline{q} - \underline{\epsilon} + \underline{\epsilon})' W(\underline{c} - \underline{q} - \underline{\epsilon} + \underline{\epsilon})\right\}^2$$

$$= E_I \left\{(\underline{c} - \underline{q} - \underline{\epsilon})' W(\underline{c} - \underline{q} - \underline{\epsilon})\right\}^2 + E\left\{2(\underline{c} - \underline{q} - \underline{\epsilon})' W\epsilon + \underline{\epsilon}'W\underline{\epsilon}\right\}^2$$

$$+ 2E_I (\underline{c} - \underline{q} - \underline{\epsilon})' W(\underline{c} - \underline{q} - \underline{\epsilon})\, \underline{\epsilon}'W\underline{\epsilon}$$

$$= (\mathrm{tr} W\Sigma)^2 + 2\mathrm{tr}(W\Sigma)^2 + 4\underline{\epsilon}'W\Sigma W\underline{\epsilon}$$

$$+ (\underline{\epsilon}'W\underline{\epsilon})^2 + 2\mathrm{tr} W\Sigma (\underline{\epsilon}'W\underline{\epsilon})$$

Hence, using (3) in the above yields

$$\mathrm{var}_I\, M = 2\mathrm{tr}(W\Sigma)^2 + 4\,\underline{\epsilon}'\, W\Sigma W\underline{\epsilon} \tag{6}$$

REFERENCES

1. Cooper, Paul W., Statistical Classification With Quadratic Forms, Biometrika, Vol. 50, Dec. (1963).
2. Anderson, T. W., An Introduction to Multivariate Analysis, Wiley, 1958.
3. Kendall, M. G., and Stuart (1958, 1961), The Advanced Theory of Statistics, 1 and 2, London, Charles Griffin and Co.

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author) Westat Research, Inc. 1395 Allison Street Denver, Colorado 80215	2a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED 2b. GROUP

3. REPORT TITLE

ASSOCIATIVE ADJUSTMENTS TO REDUCE ERRORS IN DOCUMENT SCREENING

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

Scientific Final

5. AUTHOR(S) (First name, middle initial, last name)

Bryant, Edward C.
Shumway, Robert H.
Weinman, David G.
Searls, Donald T.

6. REPORT DATE March 14, 1967	7a. TOTAL NO. OF PAGES 77	7b. NO. OF REFS 15

8a. CONTRACT OR GRANT NO AF49(638)-1671 b. PROJECT NO. 9769-02 c. 61445014 d. 681304	9a. ORIGINATOR'S REPORT NUMBER(S) 66-301 9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report) AFOSR 67-0980

10. DISTRIBUTION STATEMENT

Distribution of this document is unlimited.

11. SUPPLEMENTARY NOTES TECH, OTHER	12. SPONSORING MILITARY ACTIVITY Air Force Office of Scientific Research Directorate of Information Sciences Arlington, Virginia 22209

13. ABSTRACT

Associative adjustments to a document file have been considered as a means for improving retrieval. The investigation includes the definition and theoretical investigation of the statistical properties of a generalized mismatch measure. Improvements in retrieval resulting from performing associative regression adjustments on data file are examined both from the theoretical and experimental point of view. The expected gain in mismatch is presented as a function of various measurable characteristics of the file, such as error rates in indexing and the probability distributions of the associative adjustment criteria. Query adjustments using negative as well as positive correlations are considered and found to be ineffective. In a limited site Patent Office file with a low indexing error rate experimental results are presented applying 1) no associative correction 2) the generalized mismatch with no associative correction 3) associative correction and 4) query adjustment. In general the results using the ordinary mismatch with an associative adjustment are superior to those using the more generalized quadratic mismatch or the query adjustment scheme.

Security Classification

Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Associative transformation Associative retrieval Documentation Information retrieval Statistical association						

Security Classification